Introduction to Conformal Prediction
Extension to missing values

Margaux Zaffran

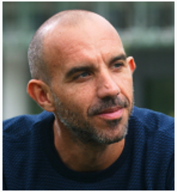
Madeleine Udell's group meeting
August 10, 2023

Who am I?

- 3rd (last) year statistics PhD Student, @ INRIA & École Polytechnique (Paris)
- Funded by Électricité de France *(French main electricity producer and supplier)*
- My advisors:

Aymeric Dieuleveut
École Polytechnique

Olivier Féron
EDF R&D
FiME



Yannig Goude
EDF R&D
LMO

Julie Josse
PreMeDICaL
INRIA

- Research interests:
  - Distribution-free uncertainty quantification
  - Time series data
  - Missing values
  - Real life applications (energy, environmental, medical and societal domains)

Conformal Prediction with Missing Values

Aymeric Dieuleveut
École Polytechnique
Paris - France

Julie Josse
PreMeDICaL
INRIA
Montpellier - France

Yaniv Romano
Technion - Israel Institute of Technology
Haifa - Israel

Setting

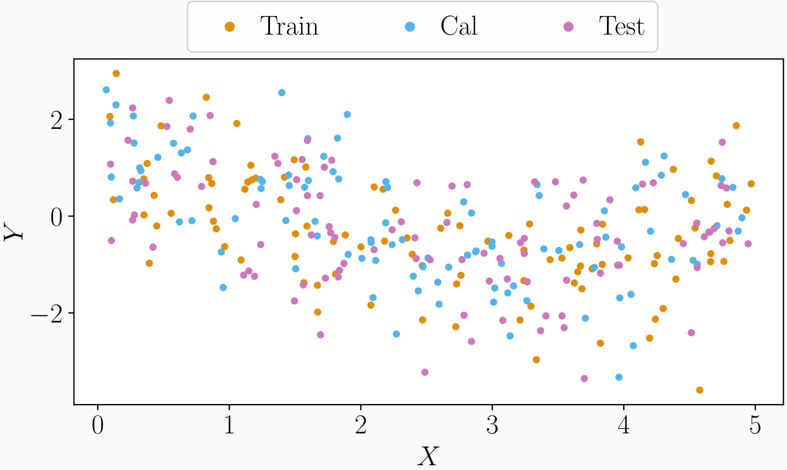
- $(X, Y) \in \mathbb{R}^d \times \mathbb{R}$ random variables
- n training samples $\left(X^{(k)}, Y^{(k)}\right)_{k=1}^{n}$
- Goal: predict an unseen point $Y^{(n+1)}$ at $X^{(n+1)}$ with **confidence**
- How? Given a miscoverage level $\alpha \in [0, 1]$, build a predictive set $\mathcal{C}_\alpha$ such that:

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

  and $\mathcal{C}_\alpha$ should be as small as possible, in order to be informative.

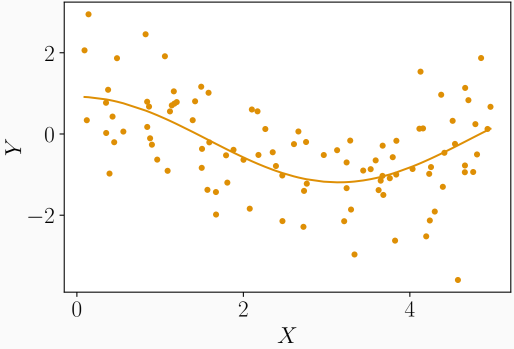
- Construction of the predictive intervals should be
  - agnostic to the model
  - agnostic to the data distribution
  - valid in finite samples

Split Conformal Prediction (SCP)[1,2,3]: toy example



[1] Vovk et al. (2005), *Algorithmic Learning in a Random World*
[2] Papadopoulos et al. (2002), *Inductive Confidence Machines for Regression*, ECML
[3] Lei et al. (2018), *Distribution-Free Predictive Inference for Regression*, JRSS B

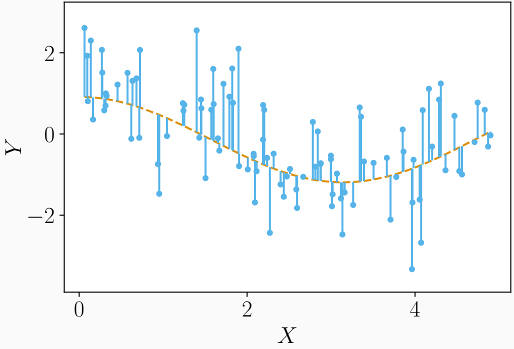
Split Conformal Prediction (SCP)[1,2,3]: training step



- Learn (or get) $\hat{\mu}$

[1] Vovk et al. (2005), *Algorithmic Learning in a Random World*
[2] Papadopoulos et al. (2002), *Inductive Confidence Machines for Regression*, ECML
[3] Lei et al. (2018), *Distribution-Free Predictive Inference for Regression*, JRSS B

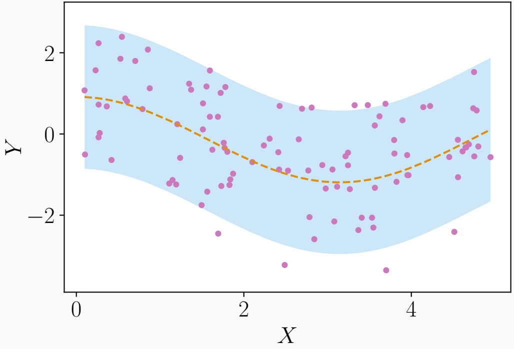
Split Conformal Prediction (SCP)[1,2,3]: calibration step



- Predict with $\hat{\mu}$
- Get the $|$residuals$|$, a.k.a. scores $\{S^{(k)}\}_{k \in \mathrm{Cal}}$
- Compute the $(1-\alpha)$ empirical quantile of $\mathcal{S} = \{|\text{residuals}|\}_{\mathrm{Cal}} \cup \{+\infty\}$, noted $q_{1-\alpha}(\mathcal{S})$

[1] Vovk et al. (2005), *Algorithmic Learning in a Random World*
[2] Papadopoulos et al. (2002), *Inductive Confidence Machines for Regression*, ECML
[3] Lei et al. (2018), *Distribution-Free Predictive Inference for Regression*, JRSS B

Split Conformal Prediction (SCP)[1,2,3]: prediction step



- Predict with $\hat{\mu}$
- Build $\widehat{C}_\alpha(x)$: $[\hat{\mu}(x) \pm q_{1-\alpha}(\mathcal{S})]$

[1] Vovk et al. (2005), *Algorithmic Learning in a Random World*
[2] Papadopoulos et al. (2002), *Inductive Confidence Machines for Regression*, ECML
[3] Lei et al. (2018), *Distribution-Free Predictive Inference for Regression*, JRSS B

SCP theoretical foundation

Definition (Exchangeability)

$\left(X^{(k)}, Y^{(k)}\right)_{k=1}^{n}$ are exchangeable if for any permutation σ of $[\![1, n]\!]$ we have:

$$\begin{aligned} & \mathcal{L}\left(\left(X^{(1)}, Y^{(1)}\right), \ldots, \left(X^{(n)}, Y^{(n)}\right)\right) \\ = \; & \mathcal{L}\left(\left(X^{(\sigma(1))}, Y^{(\sigma(1))}\right), \ldots, \left(X^{(\sigma(n))}, Y^{(\sigma(n))}\right)\right), \end{aligned}$$

where $\mathcal{L}$ designates the joint distribution.

Examples of exchangeable sequences

- i.i.d. samples
- The components of $\mathcal{N}\left(\begin{pmatrix} m \\ \vdots \\ \vdots \\ m \end{pmatrix}, \begin{pmatrix} \sigma^2 & & & \\ & \ddots & \gamma^2 & \\ & \gamma^2 & \ddots & \\ & & & \sigma^2 \end{pmatrix}\right)$

SCP: theoretical guarantees

SCP enjoys finite sample guarantees proved in Vovk et al. (2005); Lei et al. (2018).

Theorem

Suppose $\left(X^{(k)}, Y^{(k)}\right)_{k=1}^{n+1}$ *are exchangeable (or i.i.d.). SCP applied on* $\left(X^{(k)}, Y^{(k)}\right)_{k=1}^{n}$ *outputs* $\widehat{C}_\alpha(\cdot)$ *such that:*

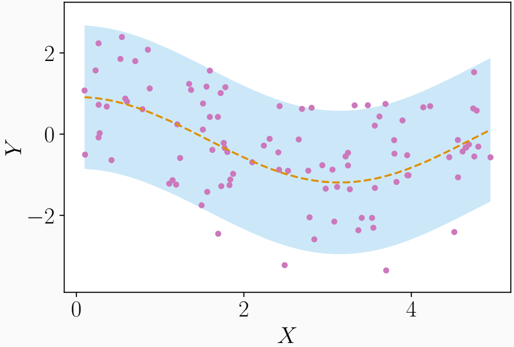
$$\mathbb{P}\left\{Y^{(n+1)} \in \widehat{C}_\alpha\left(X^{(n+1)}\right)\right\} \geq 1 - \alpha.$$

Additionally, if the scores $\left\{S^{(k)}\right\}_{k \in \mathrm{Cal}}$ *are a.s. distinct:*

$$\mathbb{P}\left\{Y^{(n+1)} \in \widehat{C}_\alpha\left(X^{(n+1)}\right)\right\} \leq 1 - \alpha + \frac{1}{\#\mathrm{Cal} + 1}.$$

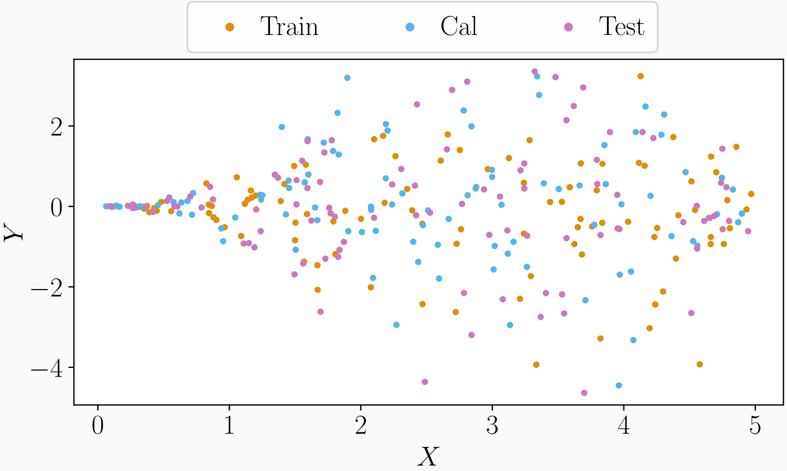
✗ Marginal coverage: $\mathbb{P}\left\{Y^{(n+1)} \in \widehat{C}_\alpha\left(X^{(n+1)}\right) \cancel{\mid X^{(n+1)} = x}\right\} \geq 1 - \alpha$

Standard mean-regression SCP is not adaptive



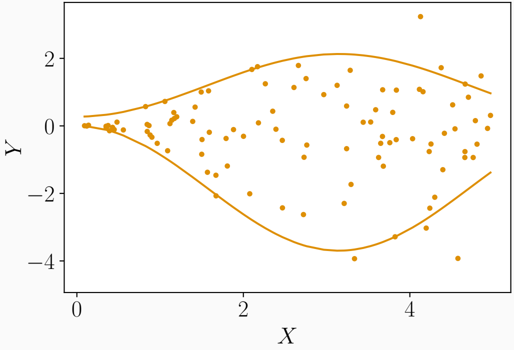
- Predict with $\hat{\mu}$
- Build $\widehat{C}_\alpha(x)$: $[\hat{\mu}(x) \pm q_{1-\alpha}(\mathcal{S})]$

Conformalized Quantile Regression (CQR)[4]



[4] Romano et al. (2019), *Conformalized Quantile Regression*, NeurIPS

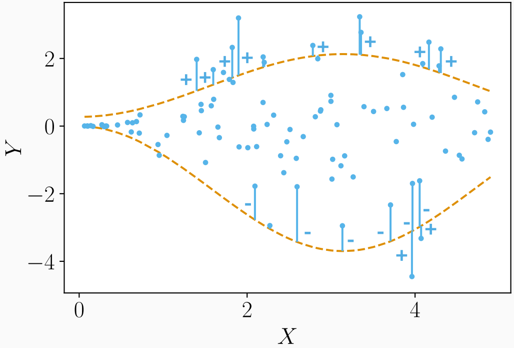
Conformalized Quantile Regression (CQR)[4]: training step



- Learn (or get) $\widehat{QR}_{\text{lower}}$ and $\widehat{QR}_{\text{upper}}$

[4] Romano et al. (2019), *Conformalized Quantile Regression*, NeurIPS

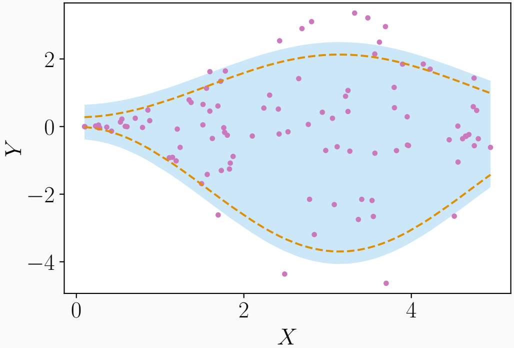
Conformalized Quantile Regression (CQR)[4]: calibration step



- Predict with $\widehat{QR}_{\text{lower}}$ and $\widehat{QR}_{\text{upper}}$
- Get the scores $\mathcal{S} = \left\{S^{(k)}\right\}_{\text{Cal}} \cup \{+\infty\}$
- Compute the $(1-\alpha)$ empirical quantile of $\mathcal{S}$, noted $q_{1-\alpha}(\mathcal{S})$

$$\hookrightarrow \quad S^{(k)} := \max\left\{\widehat{QR}_{\text{lower}}\left(X^{(k)}\right) - Y^{(k)}, Y^{(k)} - \widehat{QR}_{\text{upper}}\left(X^{(k)}\right)\right\}$$

[4] Romano et al. (2019), *Conformalized Quantile Regression*, NeurIPS



- Predict with $\widehat{QR}_{\text{lower}}$ and $\widehat{QR}_{\text{upper}}$

- Build
$$\widehat{C}_{\alpha}(x) = [\widehat{QR}_{\text{lower}}(x) - q_{1-\alpha}(\mathcal{S}); \widehat{QR}_{\text{upper}}(x) + q_{1-\alpha}(\mathcal{S})]$$

[4] Romano et al. (2019), *Conformalized Quantile Regression*, NeurIPS

Generalization: SCP is defined by the conformity scores

1. Split randomly the training data into a proper training set (size $\#\mathrm{Tr}$) and a calibration set (size $\#\mathrm{Cal}$)
2. Train your algorithm on the proper training set to obtain $\hat{A}$
3. On the calibration set, obtain $\#\mathrm{Cal}+1$ conformity scores

$$\mathcal{S} = \{S^{(k)} = \mathrm{s}\left(X^{(k)}, Y^{(k)}\right), k \in \mathrm{Cal}\} \cup \{+\infty\}$$

   Ex 1: $\mathrm{s}(x, y) = |\hat{A}(x) - y|$ in mean-regression with standard scores

   Ex 2: $\mathrm{s}(x, y) = \max\left(\widehat{QR}_{\alpha/2}(x) - y, y - \widehat{QR}_{1-\alpha/2}(x)\right)$ in CQR

4. Compute the $1-\alpha$ quantile of these scores, noted $q_{1-\alpha}(\mathcal{S})$
5. For a new point $X^{(n+1)}$, return

$$\widehat{C}_\alpha\left(X^{(n+1)}\right) := \{y \text{ such that } \mathrm{s}(\hat{A}\left(X^{(n+1)}\right), y) \leq q_{1-\alpha}(\mathcal{S})\}$$

$\hookrightarrow$ The definition of the conformity scores is crucial, as they incorporate almost all the information: data + underlying model

SCP: theoretical guarantees generalized

SCP enjoys finite sample guarantees proved in Vovk et al. (2005); Lei et al. (2018).

Theorem

Suppose $\left(X^{(k)}, Y^{(k)}\right)_{k=1}^{n+1}$ *are exchangeable (or i.i.d.). SCP applied on* $\left(X^{(k)}, Y^{(k)}\right)_{k=1}^{n}$ *outputs* $\widehat{C}_\alpha\left(X^{(n+1)}\right)$ *such that:*

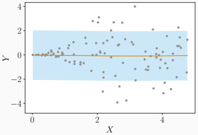
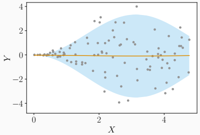
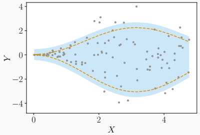
$$\mathbb{P}\left\{Y^{(n+1)} \in \widehat{C}_\alpha\left(X^{(n+1)}\right)\right\} \geq 1 - \alpha.$$

Additionally, if the scores $\left\{S^{(k)}\right\}_{k \in \mathrm{Cal}}$ *are a.s. distinct:*

$$\mathbb{P}\left\{Y^{(n+1)} \in \widehat{C}_\alpha\left(X^{(n+1)}\right)\right\} \leq 1 - \alpha + \frac{1}{\#\mathrm{Cal} + 1}.$$

✗ Marginal coverage: $\mathbb{P}\left\{Y^{(n+1)} \in \widehat{C}_\alpha\left(X^{(n+1)}\right) \cancel{| X^{(n+1)} = x}\right\} \geq 1 - \alpha$

SCP: what choices for the regression scores?

	Standard SCP Vovk et al. (2005)	**Locally weighted SCP** Lei et al. (2018)	**CQR** Romano et al. (2019)
$s(X, Y)$	$\lvert \hat{A}(X) - Y \rvert$	$\frac{\lvert \hat{A}(X) - Y \rvert}{\hat{\rho}(X)}$	$\max(\widehat{QR}_{\alpha/2}(X) - Y, Y - \widehat{QR}_{1-\alpha/2}(X))$
$\widehat{C}_\alpha(x)$	$\left[\hat{A}(x) \pm q_{1-\alpha}(\mathcal{S})\right]$	$\left[\hat{A}(x) \pm q_{1-\alpha}(\mathcal{S})\hat{\rho}(x)\right]$	$[\widehat{QR}_{\alpha/2}(x) - q_{1-\alpha}(\mathcal{S});$ $\widehat{QR}_{1-\alpha/2}(x) + q_{1-\alpha}(\mathcal{S})]$
Visu.			
✓	black-box around a "usable" prediction	black-box around a "usable" prediction	adaptive
✗	not adaptive	limited adaptiveness	no black-box around a "usable" prediction

SCP: summary

Split conformal prediction is simple to compute and works:

- any regression (and classification) algorithm (neural nets, random forest...);
- distribution-free as long as the data is exchangeable;
- finite sample.

Two interests:

- quantify the uncertainty of the underlying model $\hat{A}$;
- output predictive regions.

Note that the theoretical guarantee is **marginal** over the joint distribution of (X, Y), and **not conditional**. That is, there is no guarantee that for any $x \in \mathbb{R}$:

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

Motivation - TraumaBase®: decision support for trauma patients

- 30 hospitals
- More than 30 000 trauma patients
- 4 000 new patients per year
- 250 continuous and categorical variables
  $\hookrightarrow$ Many useful statistical tasks

Predict the level of platelets upon arrival at hospital, given 7 covariates chosen by medical doctors.

These covariates are not always observed.

Missing values: ubiquitous in data science practice

Y	X_1	X_2	X_3	X_4	X_5	X_6
22.42	0.55	0.67	0.03	0.75	0.05	0.05
8.26	0.72	0.18	0.55	0.05	0.73	0.50
~~19.41~~	~~0.60~~	~~0.58~~	~~NA~~	~~NA~~	~~NA~~	~~0.40~~
19.75	0.54	0.43	0.96	0.77	0.06	0.66
~~7.32~~	~~NA~~	~~0.19~~	~~NA~~	~~0.02~~	~~0.83~~	~~0.04~~
~~13.55~~	~~0.65~~	~~0.69~~	~~0.50~~	~~0.15~~	~~NA~~	~~0.87~~
20.75	0.43	0.74	0.61	0.72	0.52	0.35
~~9.26~~	~~0.89~~	~~NA~~	~~0.84~~	~~0.01~~	~~0.73~~	~~NA~~
~~9.68~~	~~0.963~~	~~0.45~~	~~0.65~~	~~0.04~~	~~0.06~~	~~NA~~

If each entry has a probability 0.01 of being missing:

$$d = 6 \to \ \approx 94\% \text{ of rows kept}$$

$$d = 300 \to \ \approx 5\% \text{ of rows kept}$$

One of the **ironies of Big Data** *is that missing data play an ever more significant role.*[5]

[5] Zhu et al. (2019), *High-dimensional PCA with heterogeneous missingness*, JRSS B

Handling missing values depends on pattern and mechanism

- $(X, Y) \in \mathbb{R}^d \times \mathbb{R}$ random variables.
- $M \in \{0,1\}^d$ is defined as $M_j = 1 \Leftrightarrow X_j$ is missing.
  M is called the mask or the missing pattern.

Example

We observe $(-1, \texttt{NA}, \texttt{NA})$. Then $m = (0, 1, 1)$.

There are 2^d **patterns** (statistical and computational challenges).

- Three **mechanisms**[6] can generate missing values.
  $\hookrightarrow$ **Missing Completely At Random** (MCAR): $\mathbb{P}(M = m|X) = \mathbb{P}(M = m)$ for all $m \in \{0,1\}^d$. $M \perp\!\!\!\perp X$, missingness does not depend on the variables.

[6] Rubin (1976), *Inference and missing data*, Biometrika

Supervised learning with missing values: impute-then-regress

Impute-then-regress procedures are widely used.

1. Replace NA using an imputation function ϕ (e.g. the mean).
2. Train your algorithm (Random Forest, Neural Nets, etc.) on the imputed data: $\left\{ \underbrace{\phi\left(X^{(k)}, M^{(k)}\right)}_{\text{imputed } X^{(k)}}, Y^{(k)} \right\}_{k=1}^{n}$.

$\hookrightarrow$ we consider an impute-then-regress pipeline in this work.

✓: Le Morvan et al. (2021)[7] show that for any deterministic imputation and universal learner this procedure is Bayes-consistent.

✗: Ayme et al. (2022)[8] show that even for very **simple distributions** (linear model, Gaussian noise), may suffer from **curse of dimensionality**.

[7] Le Morvan et al. (2021), *What's a good imputation to predict with missing values?*, NeurIPS

[8] Ayme et al. (2022), *Near-optimal rate of consistency for linear models with missing values*, ICML

Back to predicting the levels of platelets

- **Challenging task:** Jiang et al. (2022)[9] achieved an average relative prediction error ($\|\hat{y} - y\|^2 / \|y\|^2$) no lower than 0.23
- **Crucial task:** high-stakes decision-making problem

$\hookrightarrow$ High need for quantifying the predictive uncertainty.

[9] *Adaptive bayesian slope: Model selection with incomplete data*, Journal of Computational and Graphical Statistics

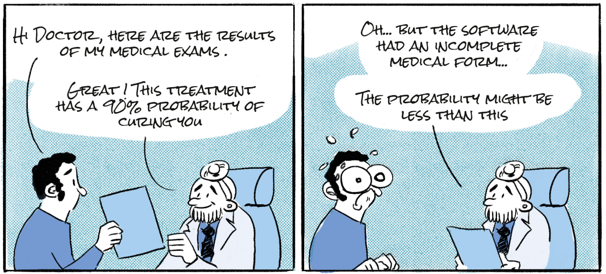
Goal: predict $Y^{(n+1)}$ with **confidence** $1-\alpha$, i.e. build the smallest $\mathcal{C}_\alpha$ such that:

1. Marginal Validity (MV)

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

2. Mask-Conditional-Validity (MCV)

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



Illustrations @theo.remlinger

CP is marginally valid (MV) after imputation

To apply conformal prediction we need **exchangeable** data.

Lemma

Assume $\left(X^{(k)}, M^{(k)}, Y^{(k)}\right)_{k=1}^{n}$ *are i.i.d. (or exchangeable).*

Then, for any missing mechanism, for almost all imputation function[10] ϕ*:*

$\left(\phi\left(X^{(k)}, M^{(k)}\right), Y^{(k)}\right)_{k=1}^{n}$ *are* ***exchangeable***.

$\Rightarrow$ CQR, and Conformal Prediction, applied on an imputed data set still enjoys marginal guarantees[11]:

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

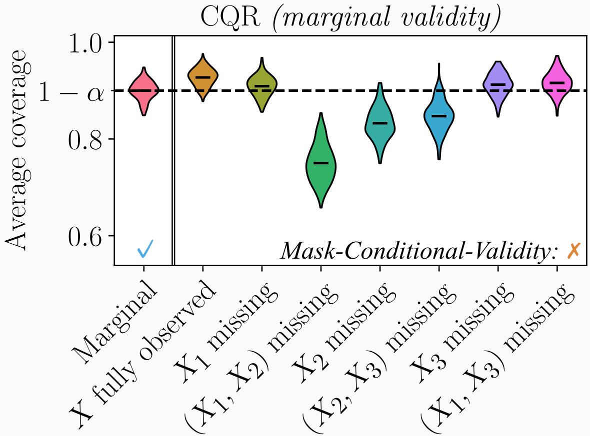
[10] Even if the imputation is not accurate, the guarantee will hold.

[11] The upper bound also holds under continuously distributed scores.

CQR is marginally valid on imputed data sets

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

with $\beta = (1, 2, -1)^T$, $\varepsilon \perp\!\!\!\perp X$, X and ε are Gaussian.



Warning: the predictive intervals cover properly **marginally**, but suffer from high **disparities depending on the missing patterns**.

Missing values induce heteroskedasticity

Theoretical study of the Gaussian linear model ($Y = \beta^T X + \varepsilon$) generalizes
$\hookrightarrow$ **oracle** intervals: smallest predictive interval when the distribution of $Y|(X, M)$ is known

Proposition (Oracle intervals under the Gaussian lin. mod.)

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

- Even with an homoskedastic noise, missingness generates heteroskedasticity
- **The uncertainty increases when missing values are associated with larger regression coefficients (i.e. the most predictive variables)**

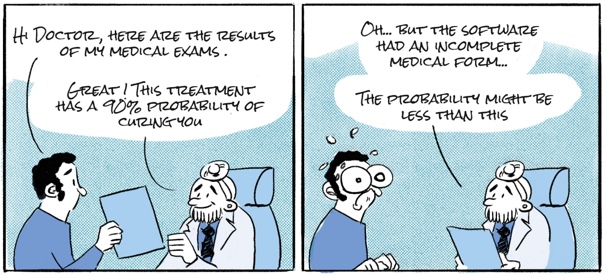
Goal: predict $Y^{(n+1)}$ with **confidence** $1-\alpha$, i.e. build the smallest $\mathcal{C}_\alpha$ such that:

1. Marginal Validity (MV) ✓

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

2. Mask-Conditional-Validity (MCV) ✗

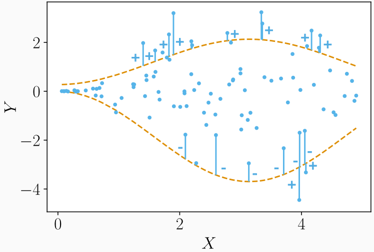
$$\forall m \in \{0,1\}^d : \mathbb{P}\left\{Y^{(n+1)} \in \mathcal{C}_\alpha\left(X^{(n+1)}, m\right) \middle| M^{(n+1)} = m\right\} \geq 1-\alpha. \tag{MCV}$$



Illustrations @theo.remlinger

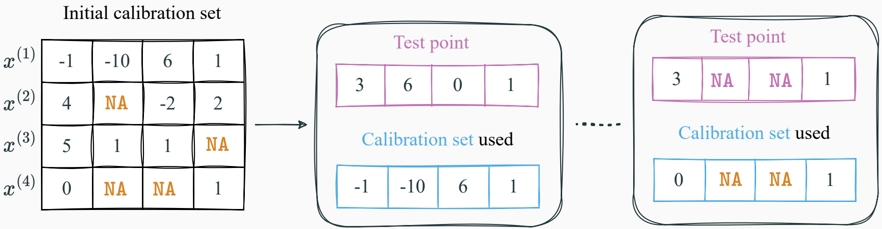
Conformalization step is independent of the important variable: the mask!

Observation: the α-correction term is computed among all the data points, regardless of their mask!



Warning: 2^d possible masks

$\Rightarrow$ Splitting the calibration set[12] by mask is infeasible (lack of data)!



[12] Romano et al. (2020), *With Malice Toward None: Assessing Uncertainty via Equalized Coverage*, Harvard Data Science Review

Missing Data Augmentation (MDA) of the calibration set

Idea: for each test point, modify the calibration points to mimic the test mask

Test point

3	NA	NA	1

Initial calibration set

$x^{(1)}$	-1	-10	6	1
$x^{(2)}$	4	NA	-2	2
$x^{(3)}$	5	1	1	NA
$x^{(4)}$	0	NA	NA	1

Calibration set used

$\tilde{x}^{(1)}$	-1	NA	NA	1
$\tilde{x}^{(2)}$	4	NA	NA	2
$\tilde{x}^{(3)}$	5	NA	NA	NA
$\tilde{x}^{(4)}$	0	NA	NA	1

Algorithms: MDA with Exact masking or with Nested masking.

CP-MDA with Exact masking

Test point

3	NA	NA	1

Initial calibration set

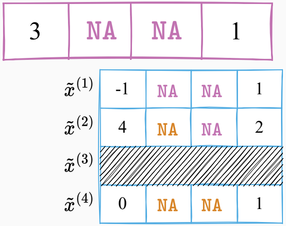
$x^{(1)}$	-1	-10	6	1
$x^{(2)}$	4	NA	-2	2
$x^{(3)}$	5	1	1	NA
$x^{(4)}$	0	NA	NA	1

Calibration set used

$\tilde{x}^{(1)}$	-1	NA	NA	1
$\tilde{x}^{(2)}$	4	NA	NA	2
$\tilde{x}^{(3)}$				
$\tilde{x}^{(4)}$	0	NA	NA	1

CQR-MDA with exact masking in words

1. Split the training set into a `proper training set` and `calibration set`
2. Train the imputation function on the `proper training set`
3. Impute the `proper training set`
4. Train the quantile regressors on the imputed `proper training set`
5. For a test point $(X^{(n+1)}, M^{(n+1)})$:



   5.1 For each $j \in [\![1, d]\!]$ s.t. $M_j^{(n+1)} = 1$, set $\tilde{M}_j^{(k)} = 1$ for k in `Cal` s.t. $M^{(k)} \subset M^{(n+1)}$
   5.2 Impute the new `calibration set`
   5.3 Compute the calibration correction, i.e. $q_{1-\alpha}(\mathcal{S})$
   5.4 Impute the test point
   5.5 Predict with the quantile regressors and the correction previously obtained, $q_{1-\alpha}(\mathcal{S})$

MDA achieves Mask-Conditional-Validity (MCV)

Theorem (CP-MDA-Exact achieves MCV)

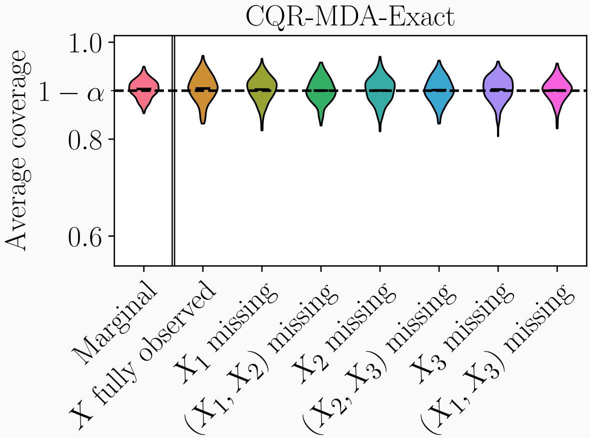
If the data is exchangeable and $M \perp\!\!\!\perp (X, Y)$, then for almost all imputation function CP-MDA-Exact is such that for any $m \in \{0,1\}^d$:

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

and if additionally the scores are almost surely distinct:

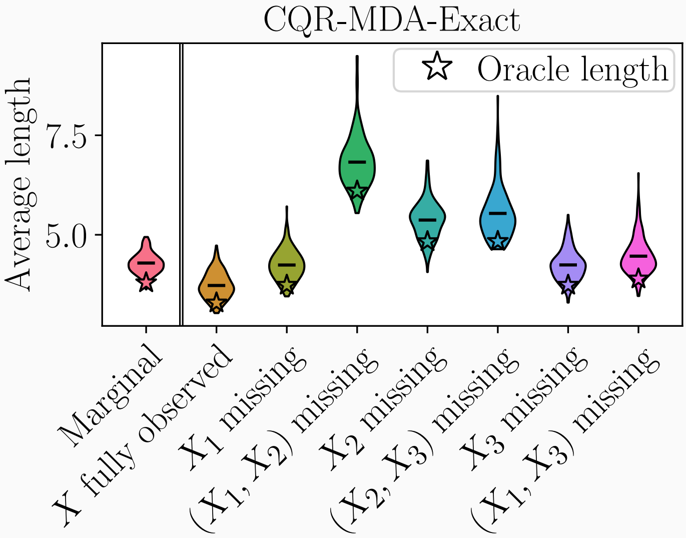
$$\mathbb{P}\left(Y \in \widehat{C}_\alpha(X, m) \,|M = m\right) \leq 1 - \alpha + \frac{1}{1 + \#\mathrm{Cal}^{\mathrm{m}}}.$$

MDA achieves Mask-Conditional-Validity (MCV), cont'd



	CQR	CQR-MDA
(MV)	✓	✓
(MCV)	✗	✓

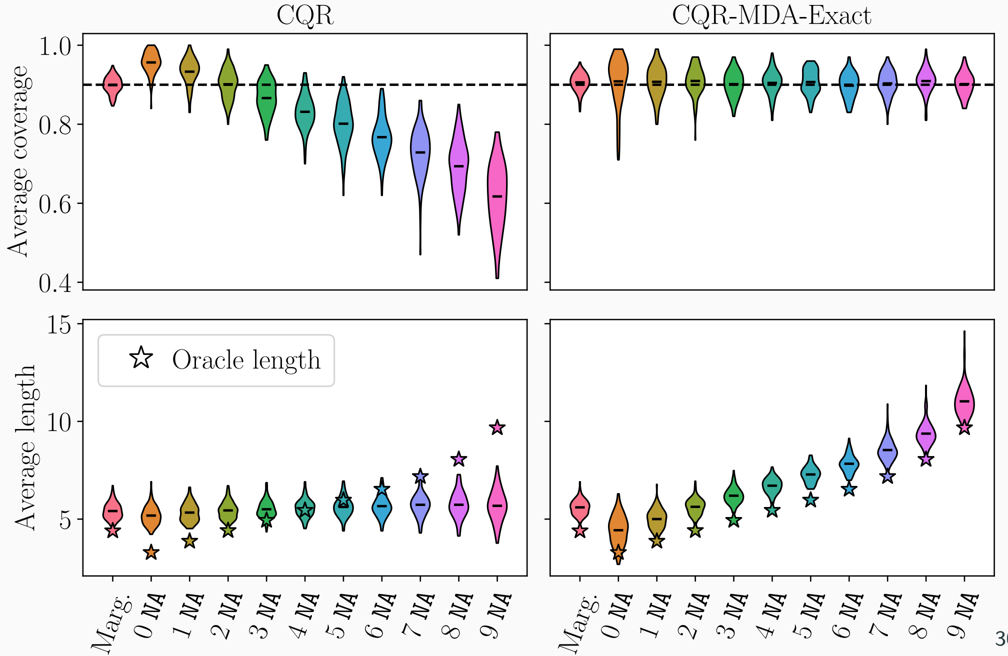
MDA achieves Mask-Conditional-Validity in an informative way



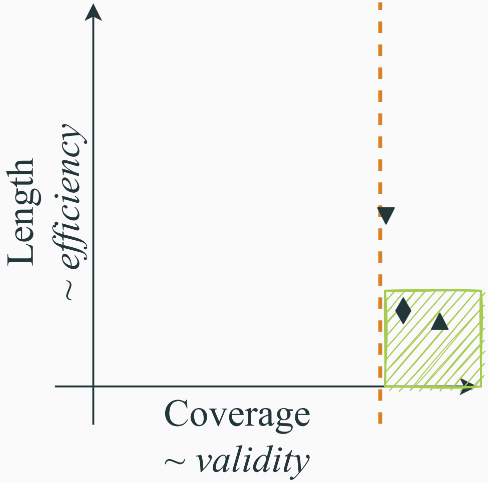
Some settings

- Imputation by iterative ridge ($\sim$ conditional expectation)
- **Concatenate the mask in the features**
- Neural network, fitted to minimize the pinball loss
- (Semi)-synthetic experiments:
  - MCAR missing values, with probability 20%
  - 100 repetitions

Synthetic experiments (Gaussian linear model, $d = 10$)



Before more experiments, visualisation



◆ : marginal coverage, i.e.

$$\mathbb{P}(Y \in \hat{C}_\alpha(X, M))$$

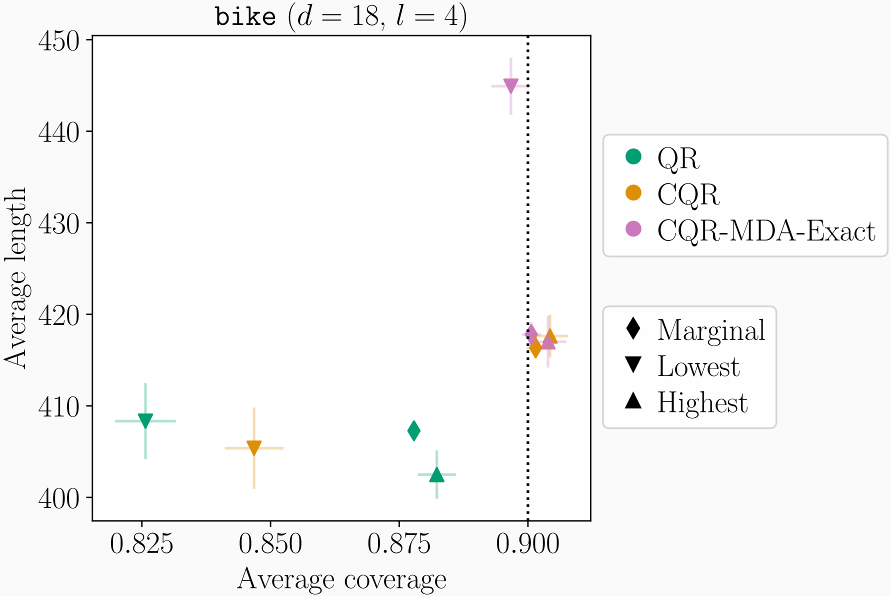
▼ : lowest coverage, i.e.

$$\min_{m \in \mathcal{M}} \mathbb{P}(Y \in \hat{C}_\alpha(X, m) | M = m)$$

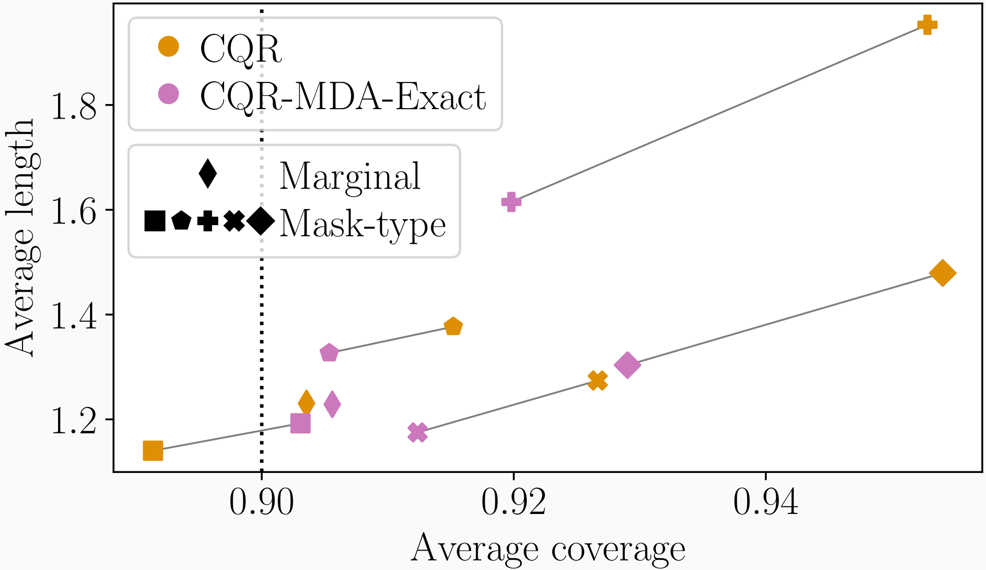
▲ : highest coverage, i.e.

$$\max_{m \in \mathcal{M}} \mathbb{P}(Y \in \hat{C}_\alpha(X, m) | M = m)$$

Semi-synthetic experiments



Real data experiment: TraumaBase®, critical care medicine



Extensions

- Consistency of universal quantile learner when chained with almost any imputation function.
- CP-MDA-Nested link to CP-MDA-Nested, an algorithm which does not discard any calibration point.

Take-home-messages

- CP marginal guarantees hold on the imputed data set.
- Missingness introduces additional heteroskedasticity, creating a need for quantile regression based methods.
- CQR fails to attain coverage conditional on the missing pattern.
- **Missing data augmentation is the first method to output predictive intervals with missing values**.
- Missing data augmentation attains conditional coverage with respect to the missing pattern (in MCAR setting).

Thank you! Questions? :)

References i

Ayme, A., Boyer, C., Dieuleveut, A., and Scornet, E. (2022). Near-optimal rate of consistency for linear models with missing values. In *ICML*.

Jiang, W., Bogdan, M., Josse, J., Majewski, S., Miasojedow, B., Ročková, V., and TraumaBase® Group (2022). Adaptive bayesian slope: Model selection with incomplete data. *Journal of Computational and Graphical Statistics*, 31(1):113–137.

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

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*.

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*, volume 119 of *Proceedings of Machine Learning Research*, pages 7130–7140. PMLR.

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

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. *NeurIPS*.

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

References iii

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

Zhu, Z., Wang, T., and Samworth, R. J. (2019). High-dimensional principal component analysis with heterogeneous missingness. arXiv.

Appendix

CP-MDA-Nested

CP-MDA-Exact reminder

Test point

3	NA	NA	1

Initial calibration set

$x^{(1)}$	-1	-10	6	1
$x^{(2)}$	4	NA	-2	2
$x^{(3)}$	5	1	1	NA
$x^{(4)}$	0	NA	NA	1

Calibration set used

$\tilde{x}^{(1)}$	-1	NA	NA	1
$\tilde{x}^{(2)}$	4	NA	NA	2
$\tilde{x}^{(3)}$				
$\tilde{x}^{(4)}$	0	NA	NA	1

What if we kept all individuals?

Test point

3	NA	NA	1

Initial calibration set

$x^{(1)}$	-1	-10	6	1
$x^{(2)}$	4	NA	-2	2
$x^{(3)}$	5	1	1	NA
$x^{(4)}$	0	NA	NA	1

Calibration set used

$\tilde{x}^{(1)}$	-1	NA	NA	1
$\tilde{x}^{(2)}$	4	NA	NA	2
$\tilde{x}^{(3)}$	5	NA	NA	NA
$\tilde{x}^{(4)}$	0	NA	NA	1

Idea: modify the test point accordingly

Test point

3	NA	NA	1

Initial calibration set

$x^{(1)}$	-1	-10	6	1
$x^{(2)}$	4	NA	-2	2
$x^{(3)}$	5	1	1	NA
$x^{(4)}$	0	NA	NA	1

Calibration set used

$\tilde{x}^{(1)}$	-1	NA	NA	1
$\tilde{x}^{(2)}$	4	NA	NA	2
$\tilde{x}^{(3)}$	5	NA	NA	NA
$\tilde{x}^{(4)}$	0	NA	NA	1

and

Temporary test points

3	NA	NA	1
3	NA	NA	1
3	NA	NA	NA
3	NA	NA	1

CQR-MDA with nested masking in words

1. For a test point $(X^{(n+1)}, M^{(n+1)})$:

3	NA	NA	1

$\tilde{x}^{(1)}$	-1	NA	NA	1
$\tilde{x}^{(2)}$	4	NA	NA	2
$\tilde{x}^{(3)}$	5	NA	NA	NA
$\tilde{x}^{(4)}$	0	NA	NA	1

   1.1 Set $\tilde{M}^{(k)} = \max(M^{(k)}, M^{(n+1)})$ for k in the `calibration set`

   1.2 Impute the new `calibration set`

   1.3 For each augmented calibration point k:

      1.3.1 Get its score $S^{(k)}$

      1.3.2 Impute-then-predict on the augmented test point $(X^{(n+1)}, \tilde{M}^{(k)})$, giving: $\widehat{QR}_{\alpha/2}(\tilde{X}^{(n+1),k})$ and $\widehat{QR}_{1-\alpha/2}(\tilde{X}^{(n+1),k})$

3	NA	NA	1
3	NA	NA	1
3	NA	NA	NA
3	NA	NA	1

      1.3.3 Compute the corrected prediction interval:
      $$[\widehat{QR}_{\alpha/2}(\tilde{X}^{(n+1),k}) - S^{(k)}; \widehat{QR}_{1-\alpha/2}(\tilde{X}^{(n+1),k}) + S^{(k)}] := \left[Z_{\text{inf}}^{(k)}; Z_{\text{sup}}^{(k)}\right]$$

   1.4 Compute the quantiles $q_\alpha(\{Z_{\text{inf}}^{(k)}\}_{k\in\mathrm{Cal}})$ and $q_{1-\alpha}(\{Z_{\text{sup}}^{(k)}\}_{k\in\mathrm{Cal}})$

   1.5 Predict $[q_\alpha(\{Z_{\text{inf}}^{(k)}\}_{k\in\mathrm{Cal}}); q_{1-\alpha}(\{Z_{\text{sup}}^{(k)}\}_{k\in\mathrm{Cal}})]$

Summary of CP-MDA

Test point

3	NA	NA	1

Initial calibration set

$x^{(1)}$	-1	-10	6	1
$x^{(2)}$	4	NA	-2	2
$x^{(3)}$	5	1	1	NA
$x^{(4)}$	0	NA	NA	1

CP-MDA with exact masking:

calibration set

$\tilde{x}^{(1)}$	-1	NA	NA	1
$\tilde{x}^{(2)}$	4	NA	NA	2
$\tilde{x}^{(3)}$				
$\tilde{x}^{(4)}$	0	NA	NA	1

CP-MDA with nested masking:

calibration set

$\tilde{x}^{(1)}$	-1	NA	NA	1
$\tilde{x}^{(2)}$	4	NA	NA	2
$\tilde{x}^{(3)}$	5	NA	NA	NA
$\tilde{x}^{(4)}$	0	NA	NA	1

and

temporary test points

3	NA	NA	1
3	NA	NA	1
3	NA	NA	NA
3	NA	NA	1

Towards asymptotic individualized coverage

Consistency of a universal quantile learner after imputation

Let Φ be an imputation function chosen by the user.

Denote $g^*_{\beta,\Phi} \in \underset{g:\mathbb{R}^d \to \mathbb{R}}{\operatorname{argmin}} \ \mathbb{E}\left[\rho_\beta(Y - g \circ \Phi(X, M))\right] := \mathcal{R}_{\beta,\phi}(g)$.

Comparison with: $\underset{f}{\operatorname{argmin}} \ \mathbb{E}\left[\rho_\beta(Y - f(X, M))\right]$ *(informal)*.

Proposition (Pinball-consistency of an universal learner)

For almost all $\mathcal{C}^\infty$ imputation function Φ, the function $g^*_{\beta,\Phi} \circ \Phi$ is Bayes optimal for the pinball-risk of level β.

$\hookrightarrow$ any universally consistent algorithm for quantile regression trained on the data imputed by Φ is pinball-Bayes-consistent.

This is an extension of the result of Le Morvan et al. (2021).

Asymptotic conditional coverage of a universal quantile learner

Corollary

For any missing mechanism, for almost all $\mathcal{C}^\infty$ imputation function Φ, if $F_{Y|(X_{\mathrm{obs(M)}},M)}$ is continuous, a universally consistent quantile regressor trained on the imputed data set yields asymptotic conditional coverage.

$\hookrightarrow \mathbb{P}(Y \in \widehat{C}_\alpha(x) | X = x, M = m) \geq 1 - \alpha$ for any $m \in \mathcal{M}$ and any $x \in \mathbb{R}^d$, asymptotically with a super quantile learner.

$$d = 3$$

Data generation

$(X, Y) \in \mathbb{R}^3 \times \mathbb{R}$.

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

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

$$(X_1, X_2, X_3) \sim \mathcal{N}\left(\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 & 0.8 & 0.8 \\ 0.8 & 1 & 0.8 \\ 0.8 & 0.8 & 1 \end{pmatrix}\right).$$

All components of X each have a probability 0.2 of being missing, Completely At Random.

Simulation settings

- Method: CQR
- Basemodel: neural network
- 200 repetitions
  - `train` size of 250 points
  - `calibration` size of 250 points
  - `test` size of 2000 points

$d = 10$, with missing data augmentation

Data generation

$(X, Y) \in \mathbb{R}^{10} \times \mathbb{R}$.

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

with $\varepsilon \sim \mathcal{N}(0,1)$, $\beta = (1, 2, -1, 3, -0.5, -1, 0.3, 1.7, 0.4, -0.3)^T$ and

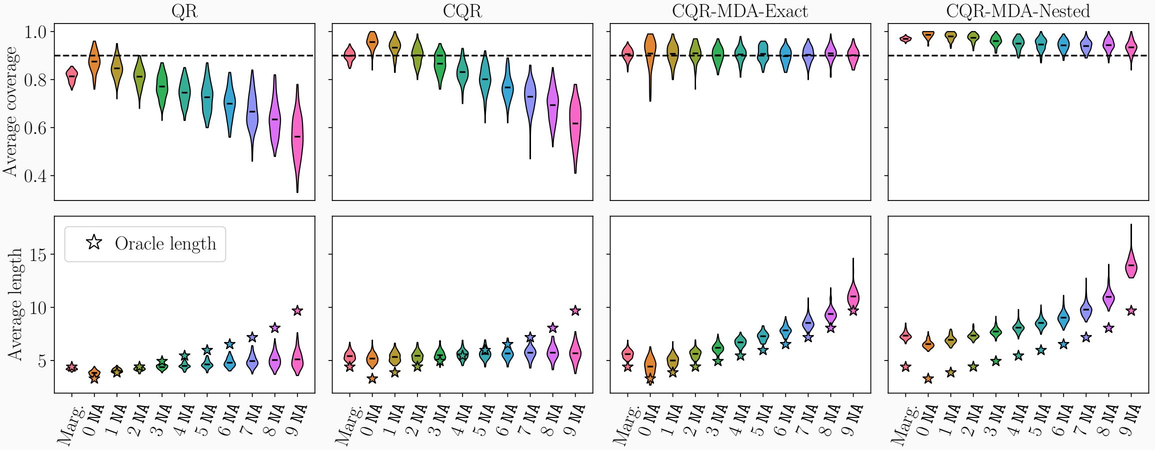
$$(X_1, \cdots, X_{10}) \sim \mathcal{N}\left(\begin{pmatrix} 1 \\ \vdots \\ \vdots \\ 1 \end{pmatrix}, \begin{pmatrix} 1 & 0.8 & \cdots & 0.8 \\ 0.8 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0.8 \\ 0.8 & \cdots & 0.8 & 1 \end{pmatrix} \right).$$

All components of X each have a probability 0.2 of being missing, Completely At Random.

Simulation settings

- Method: CQR
- Basemodel: neural network
- Imputation: iterative ($\approx$ conditional expectation)
- Mask as features: yes
- 100 repetitions
  - `train size` of 500 points
  - `calibration size` of 250 points
  - `test size` of 100 points for each pattern size, and 2000 for the marginal test set

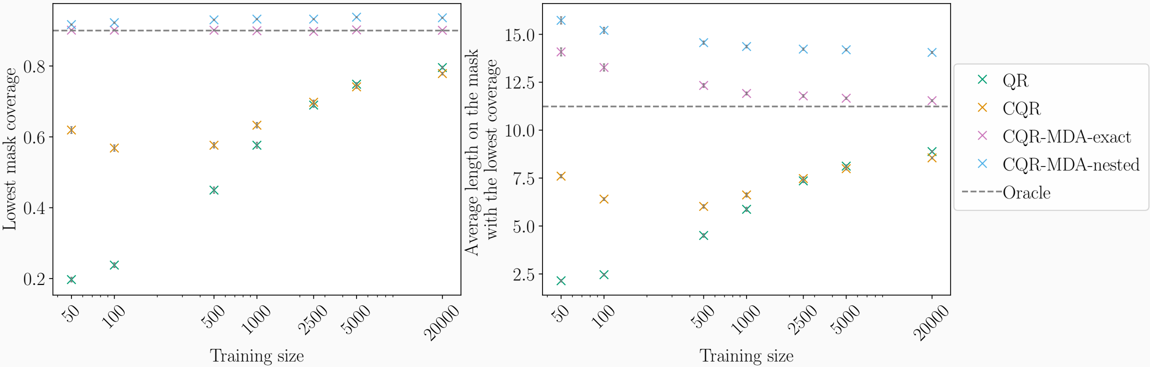
Results per pattern size



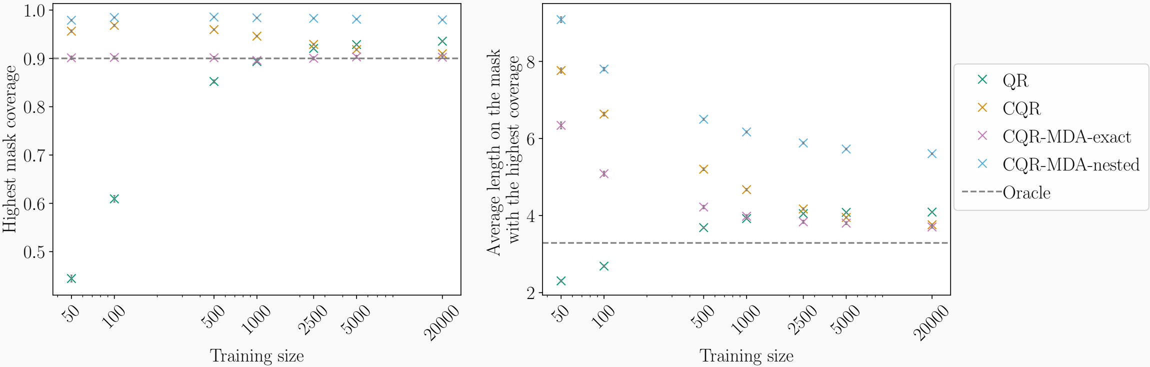
Simulation settings: varying training size

- Method: CQR
- Basemodel: neural network
- Imputation: iterative ($\approx$ conditional expectation)
- Mask as features: yes
- 100 repetitions
    - `train size` varies
    - `calibration size` of 1000 points
    - `test size` of 2000 points

Results on the worst group



Results on the best group



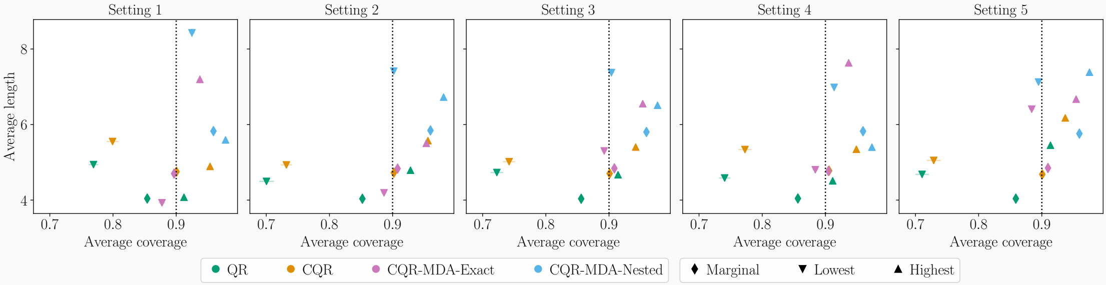
Simulation settings: beyond MCAR

- 6 variables (denote this set X_{missing}) out of 10 can be missing (the 4 others form the set X_{observed})

  $\rightarrow$ $X_{\text{missing}} = \{X_1, X_2, X_3, X_5, X_8, X_9\}$;
- Proportion of missing entries fixed to be 20%.

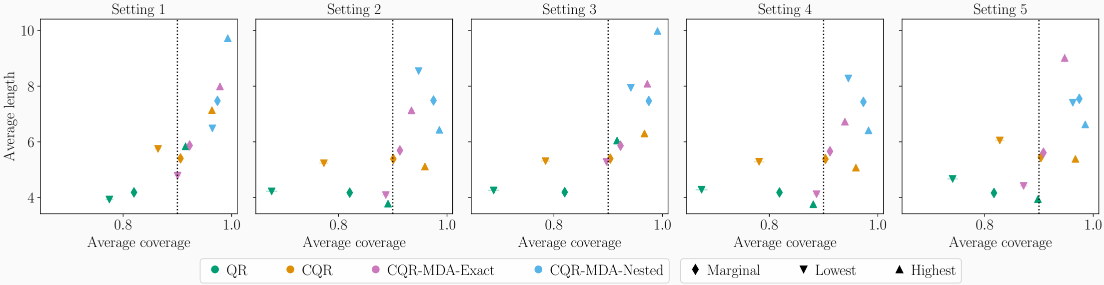
MAR missingness

- Probability of the variables in X_{missing} to be missing given by a logistic model of arguments X_{observed}.
- This setting is declined 5 times, with different weights for the logistic model.



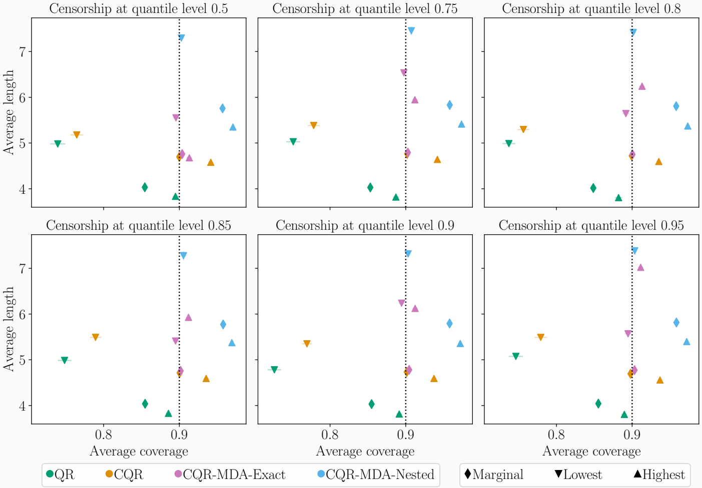
MNAR self masked missingness

- Probability of each variable in X_{missing} to be missing given by a logistic model of argument the same variable of X_{missing}.
- This setting is declined 5 times, with different weights for the logistic model.



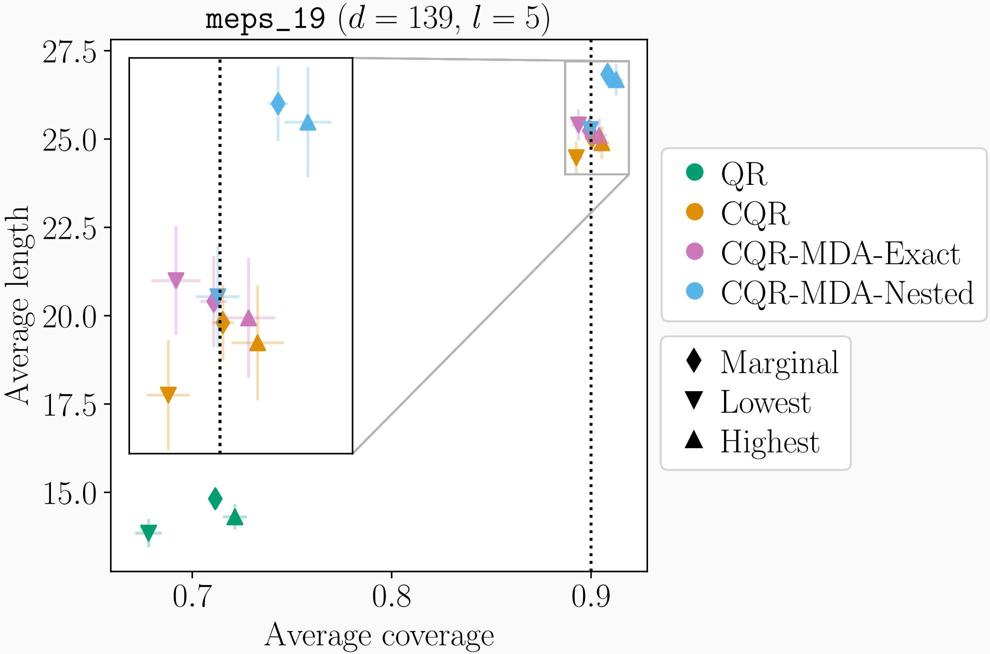
MNAR quantile censorship missingness

- Missing values are introduced at random in each q-quantile of the variables in X_{missing}.
- 6 different settings: q varies between 0.5, 0.75, 0.8, 0.85, 0.9 and 0.95.

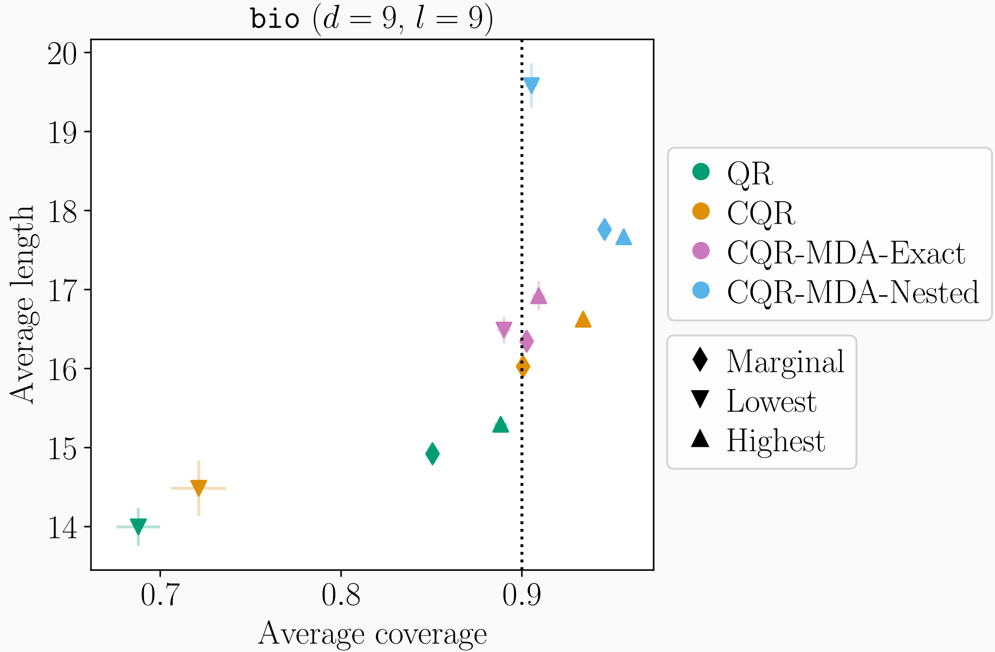


Semi-synthetic experiments with CQR-MDA-Nested

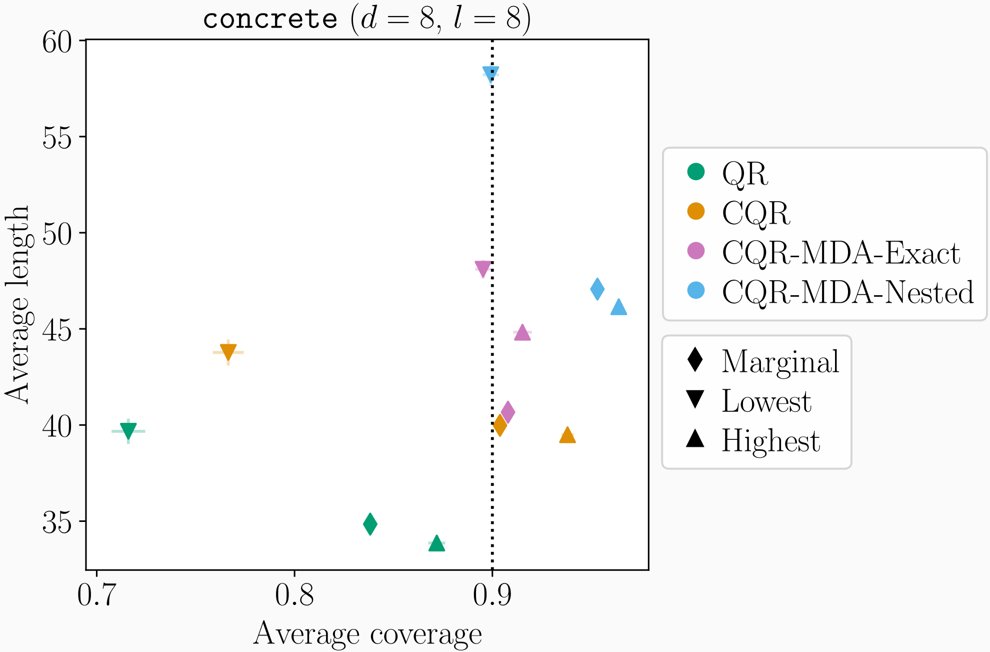
Semi-synthetic experiments with CQR-MDA-Nested



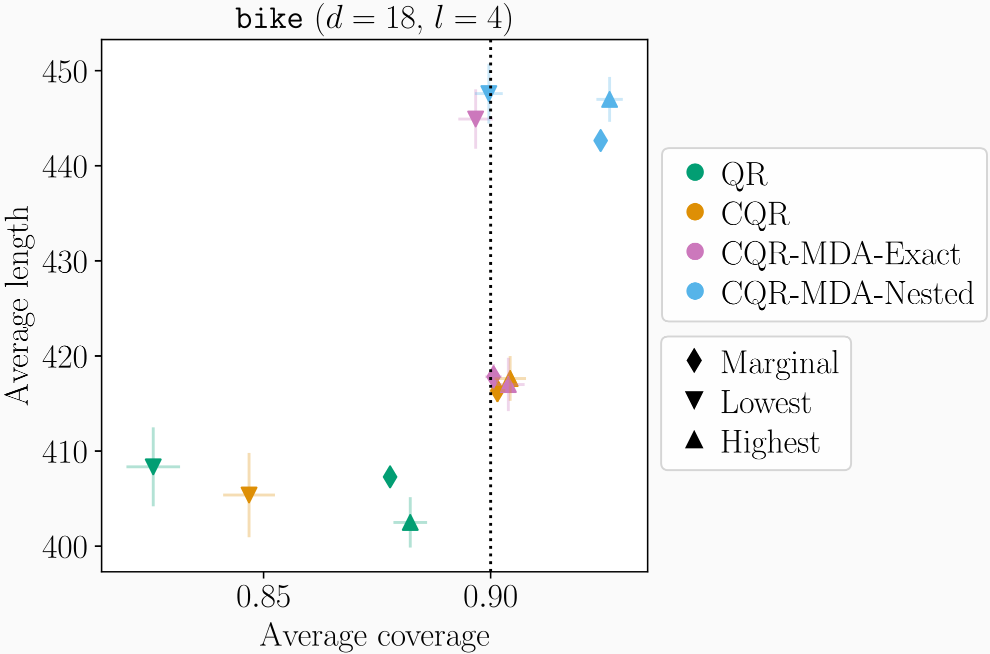
Semi-synthetic experiments with CQR-MDA-Nested



Semi-synthetic experiments with CQR-MDA-Nested



Semi-synthetic experiments with CQR-MDA-Nested



TraumaBase®

Data set description i

- `Age`: the age of the patient (no missing values);
- `Lactate`: the conjugate base of lactic acid, upon arrival at the hospital (17.66% missing values);
- `Delta_hemo`: the difference between the hemoglobin upon arrival at hospital and the one in the ambulance (23.82% missing values);
- `VE`: binary variable indicating if a Volume Expander was applied in the ambulance. A volume expander is a type of intravenous therapy that has the function of providing volume for the circulatory system (2.46% missing values);
- `RBC`: a binary index which indicates whether the transfusion of Red Blood Cells Concentrates is performed (0.37% missing values);

Data set description ii

- `SI`: the shock index. It indicates the level of occult shock based on heart rate (HR) and systolic blood pressure (SBP), that is $\mathsf{SI} = \frac{\mathsf{HR}}{\mathsf{SBP}}$, upon arrival at hospital (2.09% missing values);
- `HR`: the heart rate measured upon arrival of hospital (1.62% missing values).

Results with CQR-MDA-Nested