Conformal Prediction with Missing Values

Aymeric Dieuleveut

joint work with Margaux Zaffran, Julie Josse, Yaniv Romano

7e journée de Statistique Mathématique
January 18, 2024

Margaux Zaffrant
École Polytechnique
Inria
Paris - France

Julie Josse
PreMeDICaL
INRIA
Montpellier - France

Yaniv Romano
Technion - Israel Institute of Technology
Haifa - Israel

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

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
  ↪ Many useful statistical tasks

Predict the level of blood platelets upon arrival at hospital, given 7 pre-hospital features.

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
  ↪ Many useful statistical tasks

Predict the level of blood platelets upon arrival at hospital, given 7 pre-hospital features.

These covariates are not always observed.

Missing values: ubiquitous in data science practice

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

Y	X_1	X_2	X_3	X_4	X_5	X_6
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

Missing values: ubiquitous in data science practice

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

Y	X_1	X_2	X_3	X_4	X_5	X_6
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

Missing values: ubiquitous in data science practice

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

Y	X_1	X_2	X_3	X_4	X_5	X_6
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:

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

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

One of the **ironies of Big Data** *is that missing data play an ever more significant role.*[1]

[1] 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.

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 $(\texttt{NA}, 6, 2)$. Then $m = (1, 0, 0)$ and $X_{\text{obs}(m)} = (6, 2)$.

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 $(-1, \texttt{NA}, 2)$. Then $m = (0, 1, 0)$ and $X_{\text{obs}(m)} = (-1, 2)$.

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 $m = (0, 1, 1)$ and $X_{\text{obs}(m)} = (-1)$.

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 $m = (0, 1, 1)$ and $X_{\text{obs}(m)} = (-1)$.

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

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 $m = (0, 1, 1)$ and $X_{\text{obs}(m)} = (-1)$.

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

- Three **mechanisms**[2] can generate missing values.

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

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 $m = (0, 1, 1)$ and $X_{\text{obs}(m)} = (-1)$.

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

- Three **mechanisms**[2] can generate missing values.
  $\hookrightarrow$ **Missing Completely At Random** (MCAR): $\mathbb{P}(M = m|X) = \mathbb{P}(M = m)$ for all $m \in \{0,1\}^d$.

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

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 $m = (0, 1, 1)$ and $X_{\text{obs}(m)} = (-1)$.

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

- Three **mechanisms**[2] 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.

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

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 $m = (0, 1, 1)$ and $X_{\text{obs}(m)} = (-1)$.

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

- Three **mechanisms**[2] can generate missing values.
  ↪ Missing Completely At Random (MCAR)
  ↪ **Missing At Random** (MAR)

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

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 $m = (0, 1, 1)$ and $X_{obs(m)} = (-1)$.

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

- Three **mechanisms**[2] can generate missing values.
  ↪ Missing Completely At Random (MCAR)
  ↪ Missing At Random (MAR)
  ↪ **Missing Non At Random** (MNAR)

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

Supervised learning with missing values: impute-then-regress

Impute-then-regress procedures are widely used.

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).

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

$\xrightarrow{\phi}$

$u^{(1)}$	-1	-10	6	0
$u^{(2)}$	4	-4.5	-2	2
$u^{(3)}$	5	1	2	1
$u^{(4)}$	0	-4.5	3	1

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: $\left\{ \underbrace{\phi\left(X^{(k)}, M^{(k)}\right)}_{\text{imputed } X^{(k)}}, Y^{(k)} \right\}_{k=1}^{n}$.

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.

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)[3] show that for **any deterministic imputation** and **universal learner** this procedure is **Bayes-consistent**.

[3] Le Morvan, Josse, Scornet & Varoquaux (2021), *What's a good imputation to predict with missing values?*, NeurIPS

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)[3] show that for any deterministic imputation and universal learner this procedure is Bayes-consistent.

✗ Ayme et al. (2022)[4] show that even for very **simple distributions** (linear model, Gaussian noise), this rate of convergence may suffer from **curse of dimensionality**.

[3] Le Morvan, Josse, Scornet & Varoquaux (2021), *What's a good imputation to predict with missing values?*, NeurIPS

[4] Ayme, Boyer, Dieuleveut & Scornet (2022), *Near-optimal rate of consistency for linear models with missing values*, ICML

Predictive uncertainty quantification with missing values

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

Predictive uncertainty quantification with missing values

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

1. Marginal Validity (MV)

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

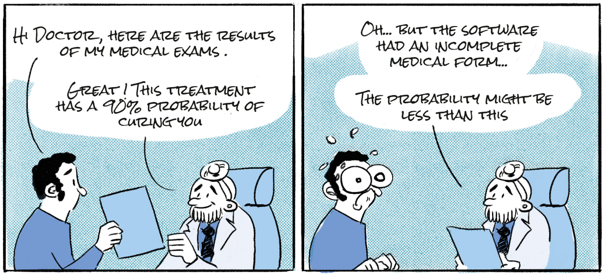
Goal: predict $Y^{(n+1)}$ with **confidence** $1-\alpha$, i.e. build the smallest $\mathcal{C}_\alpha$ such that:

1. Marginal Validity (MV)

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

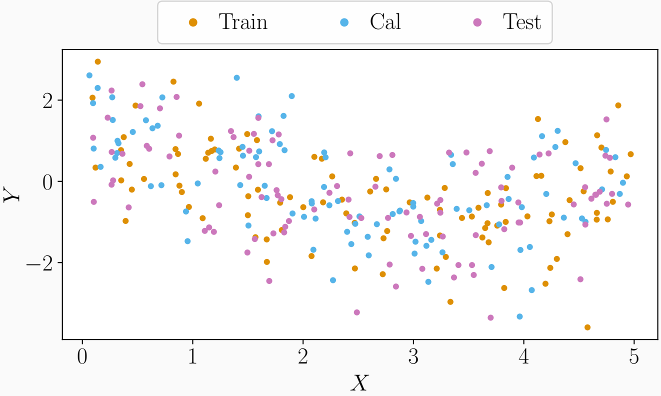
2. Mask-Conditional-Validity (MCV)

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



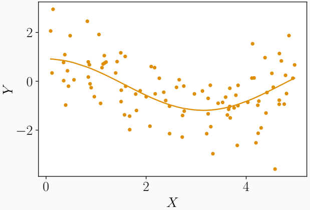
Illustrations @theo.remlinger

Split Conformal Prediction (Vovk et al., 2005): scheme



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

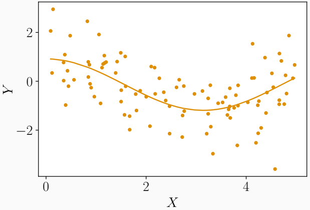
Split Conformal Prediction (Vovk et al., 2005): scheme

1) 

- Learn $\hat{\mu}$.

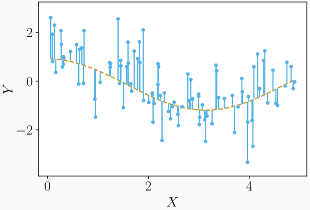
Split Conformal Prediction (Vovk et al., 2005): scheme

1)



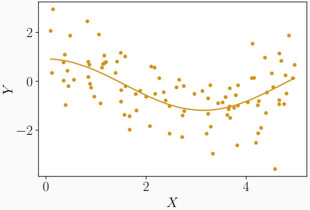
- Learn $\hat{\mu}$.

2)



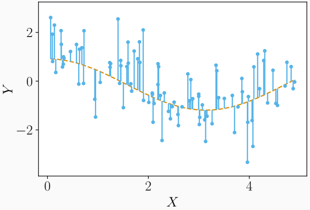
- Predict with $\hat{\mu}$.
- Get the residuals $\hat{\varepsilon}_i$ and form the set of scores $\mathcal{S} = \{|\hat{\varepsilon}_i|, i \in \mathrm{Cal}\} \cup \{+\infty\}$.
- Get their $(1 - \alpha)$ empirical quantile: $Q_{1-\alpha}(\mathcal{S})$.

1)



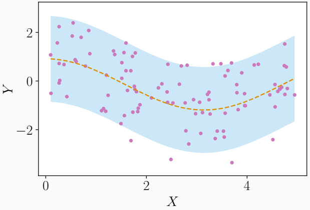
- Learn $\hat{\mu}$.

2)



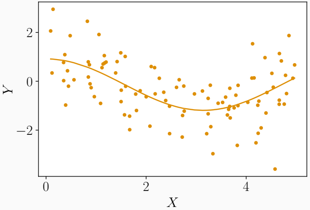
- Predict with $\hat{\mu}$.
- Get the residuals $\hat{\varepsilon}_i$ and form the set of scores $\mathcal{S} = \{|\hat{\varepsilon}_i|, i \in \mathrm{Cal}\} \cup \{+\infty\}$.
- Get their $(1-\alpha)$ empirical quantile: $Q_{1-\alpha}(\mathcal{S})$.

3)

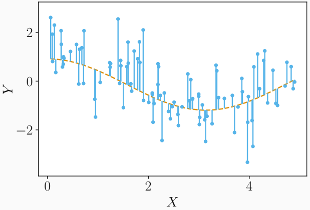


- Predict with $\hat{\mu}$.
- Build $\hat{C}_\alpha(x)$: $[\hat{\mu}(x) \pm Q_{1-\alpha}(\mathcal{S})]$.

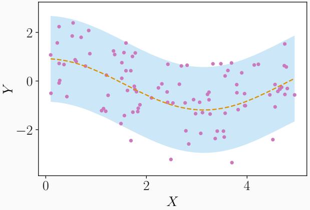
1)


2
0
−2
Y
0
2
4
X

2)


2
0
−2
Y
0
2
4
X

3)


2
0
−2
Y
0
2
4
X

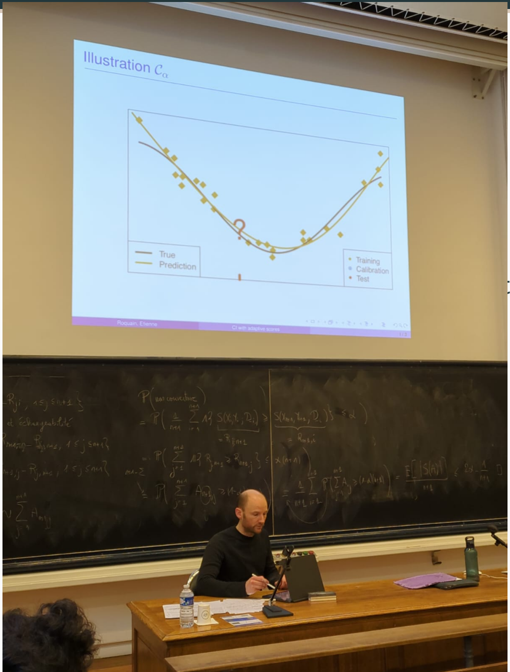
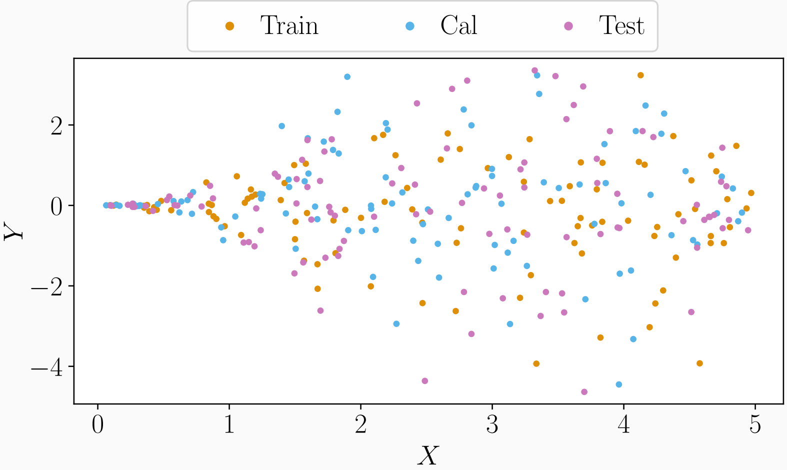


Illustration $\mathcal{C}_\alpha$
True
Prediction
Training
Calibration
Test

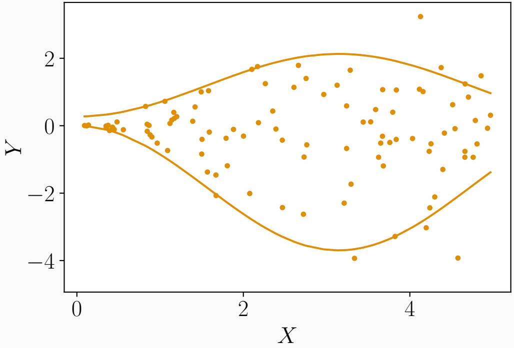
of

Conformalized Quantile Regression (CQR)[4]: toy example



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

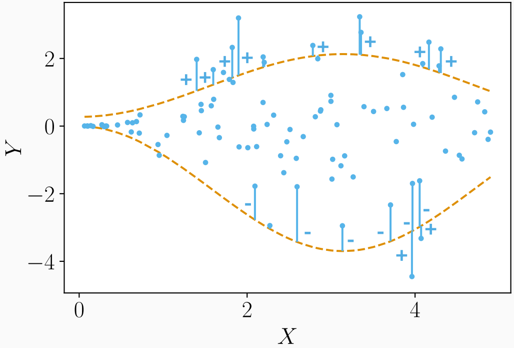
Conformalized Quantile Regression (CQR)[4]: training step



- 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

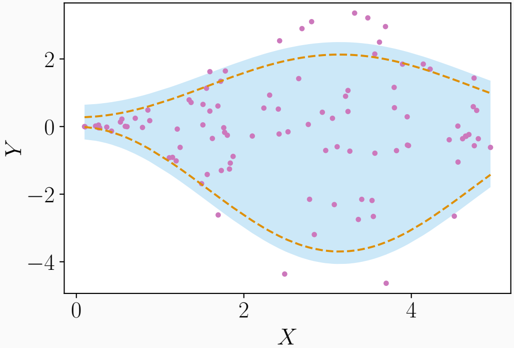


- Predict with $\widehat{QR}_{\text{lower}}$ and $\widehat{QR}_{\text{upper}}$
- Get the scores $\mathcal{S} = \left\{S^{(k)}\right\}_{\text{Cal}} \cup \{+\infty\}$
- Compute the $(1-\alpha)$ empirical quantile of $\mathcal{S}$, noted $q_{1-\alpha}(\mathcal{S})$

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

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

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



- 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

CQR: theoretical guarantees

CQR enjoys finite sample guarantees proved in Romano et al. (2019), as a particular case of Conformal Prediction (CP).

Theorem

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

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

CQR: theoretical guarantees

CQR enjoys finite sample guarantees proved in Romano et al. (2019), as a particular case of Conformal Prediction (CP).

Theorem

Suppose $\left(X^{(k)}, Y^{(k)}\right)_{k=1}^{n+1}$ are exchangeable (or i.i.d.). CQR 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:

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

✓Distribution-free, only requires exchangeability

✓Any quantile regression algorithm (neural nets, random forest...)

✓Finite sample

CQR: theoretical guarantees

CQR enjoys finite sample guarantees proved in Romano et al. (2019), as a particular case of Conformal Prediction (CP).

Theorem

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

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

✓Distribution-free, only requires exchangeability

✓Any quantile regression algorithm (neural nets, random forest...)

✓Finite sample

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

SCP is defined by the conformity score function

1. Randomly split the training data into a proper training set (size $\#\mathrm{Tr}$) and a calibration set (size $\#\mathrm{Cal}$)

SCP is defined by the conformity score function

1. Randomly split the training data into a proper training set (size $\#\mathrm{Tr}$) and a calibration set (size $\#\mathrm{Cal}$)
2. Get $\hat{A}$ *by training the algorithm* $\mathcal{A}$ *on the proper training set*

SCP is defined by the conformity score function

1. Randomly split the training data into a proper training set (size $\#\mathrm{Tr}$) and a calibration set (size $\#\mathrm{Cal}$)
2. Get $\hat{A}$ *by training the algorithm* $\mathcal{A}$ *on the proper training set*
3. On the calibration set, obtain $\#\mathrm{Cal} + 1$ conformity scores

$$\mathcal{S} = \{S_i = \mathsf{s}(\hat{A}(X_i), Y_i), i \in \mathrm{Cal}\} \cup \{+\infty\}$$

SCP is defined by the conformity score function

1. Randomly split the training data into a proper training set (size $\#\mathrm{Tr}$) and a calibration set (size $\#\mathrm{Cal}$)
2. Get $\hat{A}$ *by training the algorithm* $\mathcal{A}$ *on the proper training set*
3. On the calibration set, obtain $\#\mathrm{Cal} + 1$ conformity scores

$$\mathcal{S} = \{S_i = \mathsf{s}(\hat{A}(X_i), Y_i), i \in \mathrm{Cal}\} \cup \{+\infty\}$$

Ex 1: $\mathsf{s}(\hat{A}(X_i), Y_i) := |\hat{\mu}(X_i) - Y_i|$ in regression with standard scores

SCP is defined by the conformity score function

1. Randomly split the training data into a proper training set (size $\#\mathrm{Tr}$) and a calibration set (size $\#\mathrm{Cal}$)
2. Get $\hat{A}$ *by training the algorithm* $\mathcal{A}$ *on the proper training set*
3. On the calibration set, obtain $\#\mathrm{Cal}+1$ conformity scores

$$\mathcal{S} = \{S_i = \mathsf{s}(\hat{A}(X_i), Y_i), i \in \mathrm{Cal}\} \cup \{+\infty\}$$

Ex 1: $\mathsf{s}(\hat{A}(X_i), Y_i) := |\hat{\mu}(X_i) - Y_i|$ in regression with standard scores

Ex 2: $\mathsf{s}(\hat{A}(X_i), Y_i) := \max\left(\widehat{\mathrm{QR}}_{\mathrm{lower}}(X_i) - Y_i, Y_i - \widehat{\mathrm{QR}}_{\mathrm{upper}}(X_i)\right)$ in CQR

SCP is defined by the conformity score function

1. Randomly split the training data into a proper training set (size $\#\mathrm{Tr}$) and a calibration set (size $\#\mathrm{Cal}$)
2. Get $\hat{A}$ *by training the algorithm* $\mathcal{A}$ *on the proper training set*
3. On the calibration set, obtain $\#\mathrm{Cal}+1$ conformity scores

$$\mathcal{S} = \{S_i = \mathrm{s}(\hat{A}(X_i), Y_i), i \in \mathrm{Cal}\} \cup \{+\infty\}$$

   Ex 1: $\mathrm{s}(\hat{A}(X_i), Y_i) := |\hat{\mu}(X_i) - Y_i|$ in regression with standard scores

   Ex 2: $\mathrm{s}(\hat{A}(X_i), Y_i) := \max\left(\widehat{\mathrm{QR}}_{\mathrm{lower}}(X_i) - Y_i, Y_i - \widehat{\mathrm{QR}}_{\mathrm{upper}}(X_i)\right)$ in CQR

4. Compute the $1-\alpha$ quantile of these scores, noted $q_{1-\alpha}(\mathcal{S})$

SCP is defined by the conformity score function

1. Randomly split the training data into a proper training set (size $\#\mathrm{Tr}$) and a calibration set (size $\#\mathrm{Cal}$)
2. Get $\hat{A}$ *by training the algorithm* $\mathcal{A}$ *on the proper training set*
3. On the calibration set, obtain $\#\mathrm{Cal}+1$ conformity scores

$$\mathcal{S} = \{S_i = \mathrm{s}(\hat{A}(X_i), Y_i), i \in \mathrm{Cal}\} \cup \{+\infty\}$$

Ex 1: $\mathrm{s}(\hat{A}(X_i), Y_i) := |\hat{\mu}(X_i) - Y_i|$ in regression with standard scores

Ex 2: $\mathrm{s}(\hat{A}(X_i), Y_i) := \max\left(\widehat{\mathrm{QR}}_{\mathrm{lower}}(X_i) - Y_i, Y_i - \widehat{\mathrm{QR}}_{\mathrm{upper}}(X_i)\right)$ 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

$$\widehat{C}_\alpha(X_{n+1}) = \{y \text{ such that } \mathrm{s}(\hat{A}(X_{n+1}), y) \le q_{1-\alpha}(\mathcal{S})\}$$

SCP is defined by the conformity score function

1. Randomly split the training data into a proper training set (size $\#\mathrm{Tr}$) and a calibration set (size $\#\mathrm{Cal}$)
2. Get $\hat{A}$ *by training the algorithm* $\mathcal{A}$ *on the proper training set*
3. On the calibration set, obtain $\#\mathrm{Cal} + 1$ conformity scores

$$\mathcal{S} = \{S_i = \mathrm{s}(\hat{A}(X_i), Y_i), i \in \mathrm{Cal}\} \cup \{+\infty\}$$

   Ex 1: $\mathrm{s}(\hat{A}(X_i), Y_i) := |\hat{\mu}(X_i) - Y_i|$ in regression with standard scores

   Ex 2: $\mathrm{s}(\hat{A}(X_i), Y_i) := \max\left(\widehat{\mathrm{QR}}_{\mathrm{lower}}(X_i) - Y_i, Y_i - \widehat{\mathrm{QR}}_{\mathrm{upper}}(X_i)\right)$ 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

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

Ex 1: $\widehat{C}_{\alpha}(X_{n+1}) = [\hat{\mu}(X_{n+1}) \pm q_{1-\alpha}(\mathcal{S})]$

SCP is defined by the conformity score function

1. Randomly split the training data into a proper training set (size $\#\mathrm{Tr}$) and a calibration set (size $\#\mathrm{Cal}$)
2. Get $\hat{A}$ *by training the algorithm* $\mathcal{A}$ *on the proper training set*
3. On the calibration set, obtain $\#\mathrm{Cal}+1$ conformity scores

$$\mathcal{S} = \{S_i = \mathrm{s}(\hat{A}(X_i), Y_i), i \in \mathrm{Cal}\} \cup \{+\infty\}$$

Ex 1: $\mathrm{s}(\hat{A}(X_i), Y_i) := |\hat{\mu}(X_i) - Y_i|$ in regression with standard scores

Ex 2: $\mathrm{s}(\hat{A}(X_i), Y_i) := \max\left(\widehat{\mathrm{QR}}_{\mathrm{lower}}(X_i) - Y_i, Y_i - \widehat{\mathrm{QR}}_{\mathrm{upper}}(X_i)\right)$ 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

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

Ex 2: $\widehat{C}_\alpha(X_{n+1}) = [\widehat{QR}_{\mathrm{lower}}(X_{n+1}) - q_{1-\alpha}(\mathcal{S});$

$\widehat{QR}_{\mathrm{upper}}(X_{n+1}) + q_{1-\alpha}(\mathcal{S})]$

SCP is defined by the conformity score function

1. Randomly split the training data into a proper training set (size $\#\mathrm{Tr}$) and a calibration set (size $\#\mathrm{Cal}$)
2. Get $\hat{A}$ *by training the algorithm $\mathcal{A}$ on the proper training set*
3. On the calibration set, obtain $\#\mathrm{Cal}+1$ conformity scores

$$\mathcal{S} = \{S_i = \mathrm{s}(\hat{A}(X_i), Y_i), i \in \mathrm{Cal}\} \cup \{+\infty\}$$

Ex 1: $\mathrm{s}(\hat{A}(X_i), Y_i) := |\hat{\mu}(X_i) - Y_i|$ in regression with standard scores

Ex 2: $\mathrm{s}(\hat{A}(X_i), Y_i) := \max\left(\widehat{\mathrm{QR}}_{\mathrm{lower}}(X_i) - Y_i, Y_i - \widehat{\mathrm{QR}}_{\mathrm{upper}}(X_i)\right)$ 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

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

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

Bonus - SCP: theoretical guarantees

This procedure enjoys the finite sample guarantee proposed and proved in Vovk et al. (2005).

Theorem

Suppose $(X_i, Y_i)_{i=1}^{n+1}$ *are exchangeable*[5]. *SCP on* $(X_i, Y_i)_{i=1}^{n}$ *outputs* $\widehat{C}_\alpha(\cdot)$ *such that:*

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

If, in addition, the scores $\{S_i\}_{i \in \mathrm{Cal}} \cup \{S_{n+1}\}$ *are almost surely distinct, then*

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

Proof: application of the quantile lemma.

[5] Only the calibration and test data need to be exchangeable.

Bonus - SCP: theoretical guarantees

This procedure enjoys the finite sample guarantee proposed and proved in Vovk et al. (2005).

Theorem

Suppose $(X_i, Y_i)_{i=1}^{n+1}$ *are exchangeable*[5]. *SCP on* $(X_i, Y_i)_{i=1}^{n}$ *outputs* $\widehat{C}_\alpha(\cdot)$ *such that:*

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

If, in addition, the scores $\{S_i\}_{i \in \mathrm{Cal}} \cup \{S_{n+1}\}$ *are almost surely distinct, then*

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

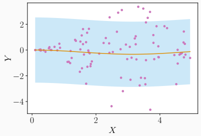
Proof: application of the quantile lemma.

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

[5] Only the calibration and test data need to be exchangeable.

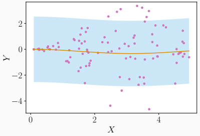
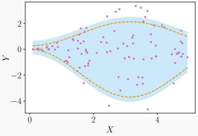
SCP: what choices for the regression scores?

$$\widehat{C}_\alpha(X_{n+1}) = \{y \text{ such that } s(\hat{A}(X_{n+1}), y) \leq q_{1-\alpha}(\mathcal{S})\}$$

	Standard SCP Vovk et al. (2005)		
$s(\hat{A}(X), Y)$	$\lvert\hat{\mu}(X) - Y\rvert$		
$\widehat{C}_\alpha(x)$	$[\hat{\mu}(x) \pm q_{1-\alpha}(\mathcal{S})]$		
Visu.			
✓	black-box around a "usable" prediction		
✗	not adaptive		

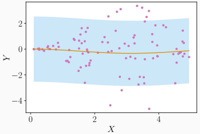
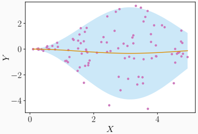
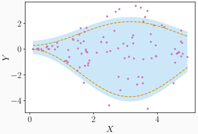
SCP: what choices for the regression scores?

$$\widehat{C}_\alpha(X_{n+1}) = \{y \text{ such that } s(\hat{A}(X_{n+1}), y) \leq q_{1-\alpha}(\mathcal{S})\}$$

	Standard SCP Vovk et al. (2005)		**CQR** Romano et al. (2019)
$s(\hat{A}(X), Y)$	$\lvert\hat{\mu}(X) - Y\rvert$		$\max(\widehat{\mathrm{QR}}_{\mathrm{lower}}(X) - Y, Y - \widehat{\mathrm{QR}}_{\mathrm{upper}}(X))$
$\widehat{C}_\alpha(x)$	$[\hat{\mu}(x) \pm q_{1-\alpha}(\mathcal{S})]$		$[\widehat{\mathrm{QR}}_{\mathrm{lower}}(x) - q_{1-\alpha}(\mathcal{S}); \widehat{\mathrm{QR}}_{\mathrm{upper}}(x) + q_{1-\alpha}(\mathcal{S})]$
Visu.			
✓	black-box around a "usable" prediction		adaptive
✗	not adaptive		no black-box around a "usable" prediction

SCP: what choices for the regression scores?

$$\widehat{C}_\alpha(X_{n+1}) = \{y \text{ such that } s(\hat{A}(X_{n+1}), y) \leq q_{1-\alpha}(\mathcal{S})\}$$

	Standard SCP Vovk et al. (2005)	**Locally weighted SCP** Lei et al. (2018)	**CQR** Romano et al. (2019)
$s(\hat{A}(X), Y)$	$\lvert\hat{\mu}(X) - Y\rvert$	$\frac{\lvert\hat{\mu}(X)-Y\rvert}{\hat{\rho}(X)}$	$\max(\widehat{\text{QR}}_{\text{lower}}(X) - Y, Y - \widehat{\text{QR}}_{\text{upper}}(X))$
$\widehat{C}_\alpha(x)$	$[\hat{\mu}(x) \pm q_{1-\alpha}(\mathcal{S})]$	$[\hat{\mu}(x) \pm q_{1-\alpha}(\mathcal{S})\hat{\rho}(x)]$	$[\widehat{\text{QR}}_{\text{lower}}(x) - q_{1-\alpha}(\mathcal{S}); \widehat{\text{QR}}_{\text{upper}}(x) + q_{1-\alpha}(\mathcal{S})]$
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

CP is marginally valid (MV) after imputation

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

Lemma (Z. et al. (2023a))

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[6] ϕ*:*

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

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

CP is marginally valid (MV) after imputation

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

Lemma (Z. et al. (2023a))

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[6] 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[7]:

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

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

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

CP is marginally valid (MV) after imputation

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

Lemma (Z. et al. (2023a))

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[6] 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[7]:

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

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

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

CP is marginally valid (MV) after imputation

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

Lemma (Z. et al. (2023a))

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[6] 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[7]:

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

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

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

CQR is marginally valid on imputed data sets

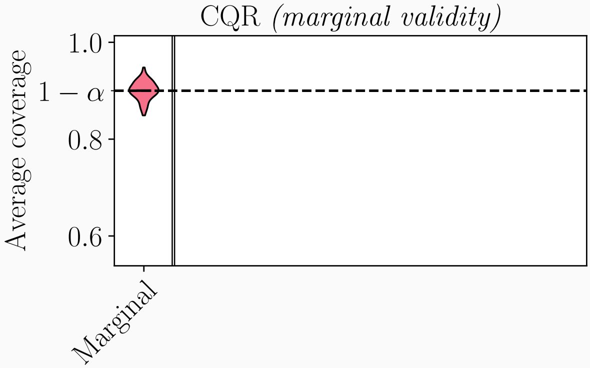
$$Y = \beta^T X + \varepsilon,$$

$\beta = (1, 2, -1)^T$, $\varepsilon \perp\!\!\!\perp X$, X and ε Gaussian, 20% uniform MCAR missing values.

CQR is marginally valid on imputed data sets

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

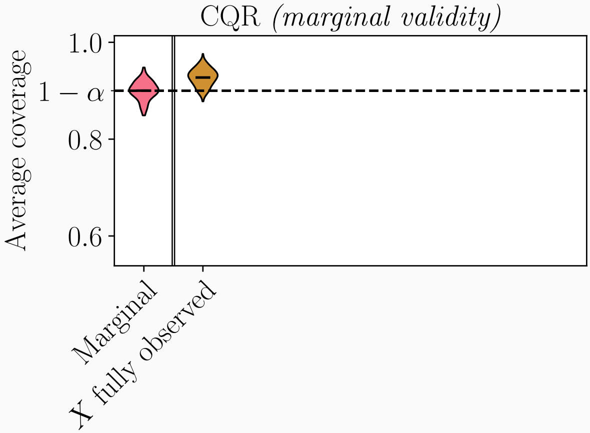
$\beta = (1, 2, -1)^T$, $\varepsilon \perp\!\!\!\perp X$, X and ε Gaussian, 20% uniform MCAR missing values.



CQR is marginally valid on imputed data sets

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

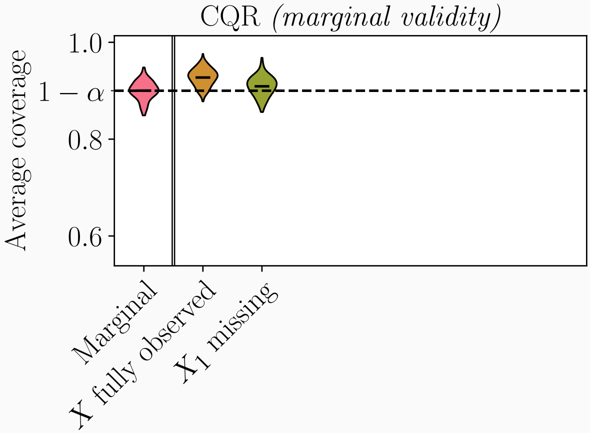
$\beta = (1, 2, -1)^T$, $\varepsilon \perp\!\!\!\perp X$, X and ε Gaussian, 20% uniform MCAR missing values.



CQR is marginally valid on imputed data sets

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

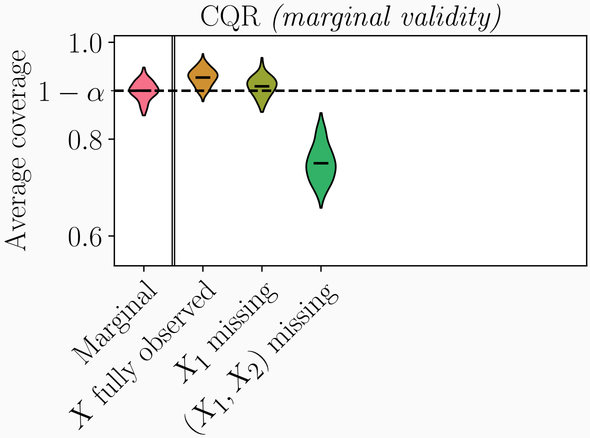
$\beta = (1, 2, -1)^T$, $\varepsilon \perp\!\!\!\perp X$, X and ε Gaussian, 20% uniform MCAR missing values.



CQR is marginally valid on imputed data sets

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

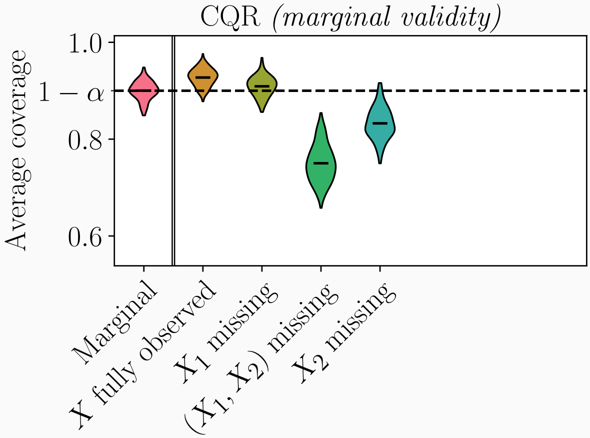
$\beta = (1, 2, -1)^T$, $\varepsilon \perp\!\!\!\perp X$, X and ε Gaussian, 20% uniform MCAR missing values.



CQR is marginally valid on imputed data sets

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

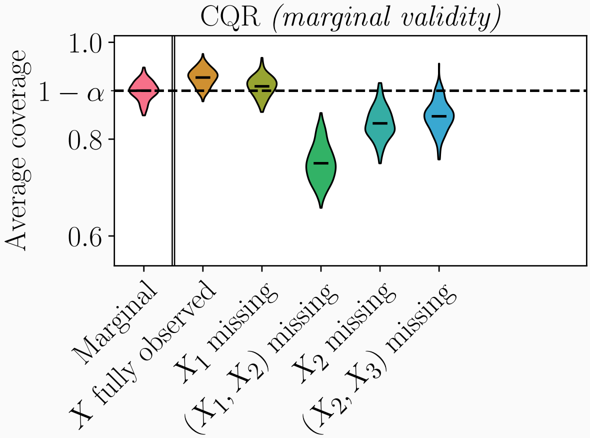
$\beta = (1, 2, -1)^T$, $\varepsilon \perp\!\!\!\perp X$, X and ε Gaussian, 20% uniform MCAR missing values.



CQR is marginally valid on imputed data sets

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

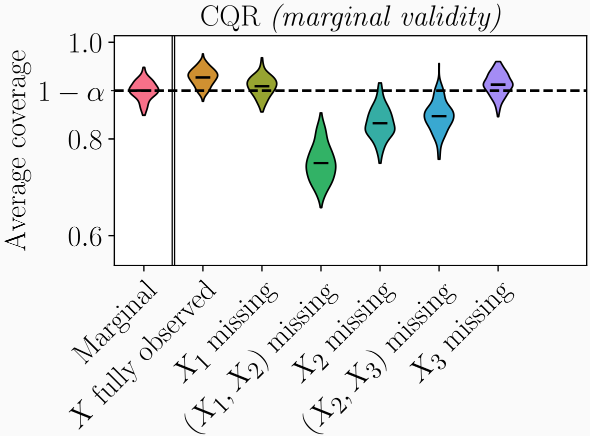
$\beta = (1, 2, -1)^T$, $\varepsilon \perp\!\!\!\perp X$, X and ε Gaussian, 20% uniform MCAR missing values.



CQR is marginally valid on imputed data sets

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

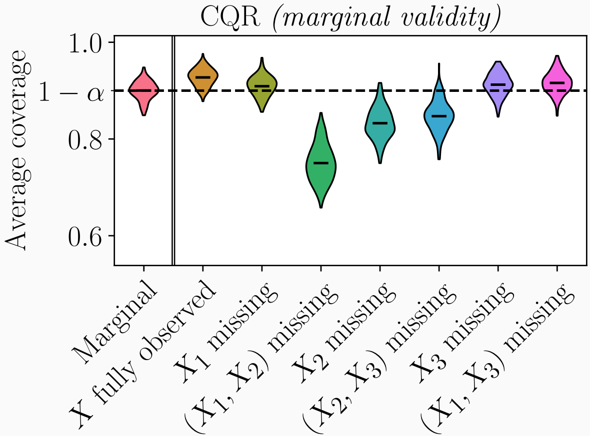
$\beta = (1, 2, -1)^T$, $\varepsilon \perp\!\!\!\perp X$, X and ε Gaussian, 20% uniform MCAR missing values.



CQR is marginally valid on imputed data sets

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

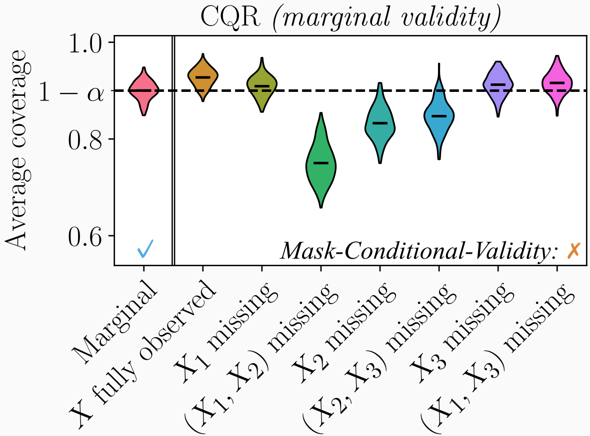
$\beta = (1, 2, -1)^T$, $\varepsilon \perp\!\!\!\perp X$, X and ε Gaussian, 20% uniform MCAR missing values.



CQR is marginally valid on imputed data sets

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

$\beta = (1, 2, -1)^T$, $\varepsilon \perp\!\!\!\perp X$, X and ε Gaussian, 20% uniform MCAR missing values.



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

Missing values induce heteroskedasticity

Gaussian linear model

- $Y = \beta^T X + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2) \perp\!\!\!\perp (X, M)$, $\beta \in \mathbb{R}^d$.
- for all $m \in \{0,1\}^d$, there exist μ^m and Σ^m such that $X|(M = m) \sim \mathcal{N}(\mu^m, \Sigma^m)$.

$\hookrightarrow$ **oracle** intervals: smallest predictive interval when the distribution of $Y|(X, M)$ is known

Missing values induce heteroskedasticity

Gaussian linear model

- $Y = \beta^T X + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2) \perp\!\!\!\perp (X, M)$, $\beta \in \mathbb{R}^d$.
- for all $m \in \{0,1\}^d$, there exist μ^m and Σ^m such that $X|(M = m) \sim \mathcal{N}(\mu^m, \Sigma^m)$.

$\hookrightarrow$ **oracle** intervals: smallest predictive interval when the distribution of $Y|(X, M)$ is known

Proposition (Oracle int. under Gaussian lin. mod., Z. et al. (2023a))

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

Missing values induce heteroskedasticity

Gaussian linear model

- $Y = \beta^T X + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2) \perp\!\!\!\perp (X, M)$, $\beta \in \mathbb{R}^d$.
- for all $m \in \{0,1\}^d$, there exist μ^m and Σ^m such that $X|(M = m) \sim \mathcal{N}(\mu^m, \Sigma^m)$.

$\hookrightarrow$ **oracle** intervals: smallest predictive interval when the distribution of $Y|(X, M)$ is known

Proposition (Oracle int. under Gaussian lin. mod., Z. et al. (2023a))

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

- Even with an homoskedastic noise, missingness generates heteroskedasticity

Missing values induce heteroskedasticity

Gaussian linear model

- $Y = \beta^T X + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2) \perp\!\!\!\perp (X, M)$, $\beta \in \mathbb{R}^d$.
- for all $m \in \{0,1\}^d$, there exist μ^m and Σ^m such that $X|(M = m) \sim \mathcal{N}(\mu^m, \Sigma^m)$.

$\hookrightarrow$ **oracle** intervals: smallest predictive interval when the distribution of $Y|(X, M)$ is known

Proposition (Oracle int. under Gaussian lin. mod., Z. et al. (2023a))

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

- 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)**

Goals reminder: achieve MCV!

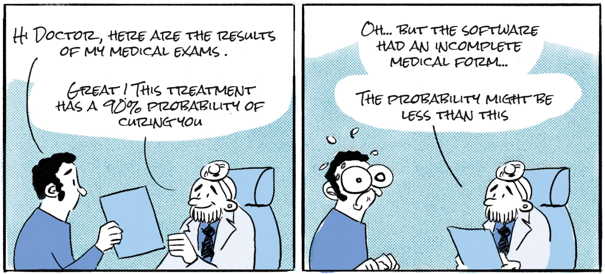
Goal: predict $Y^{(n+1)}$ with **confidence** $1-\alpha$, i.e. build the smallest $\mathcal{C}_\alpha$ such that:

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) ✗

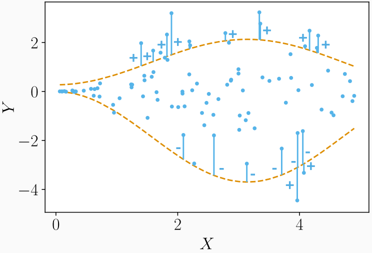
$$\forall m \in \{0,1\}^d : \mathbb{P}\left\{Y^{(n+1)} \in \mathcal{C}_\alpha\left(X^{(n+1)}, m\right) | M^{(n+1)} = m\right\} \geq 1-\alpha. \quad \text{(MCV)}$$



Illustrations @theo.remlinger

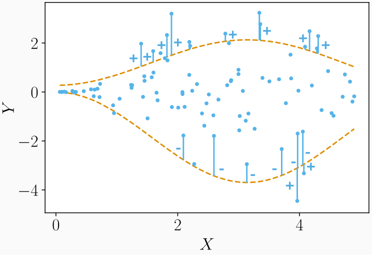
Conformalization step is independent of the important variable: the mask!

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



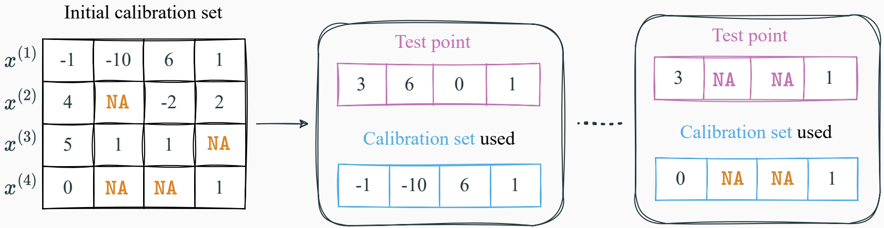
Conformalization step is independent of the important variable: the mask!

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

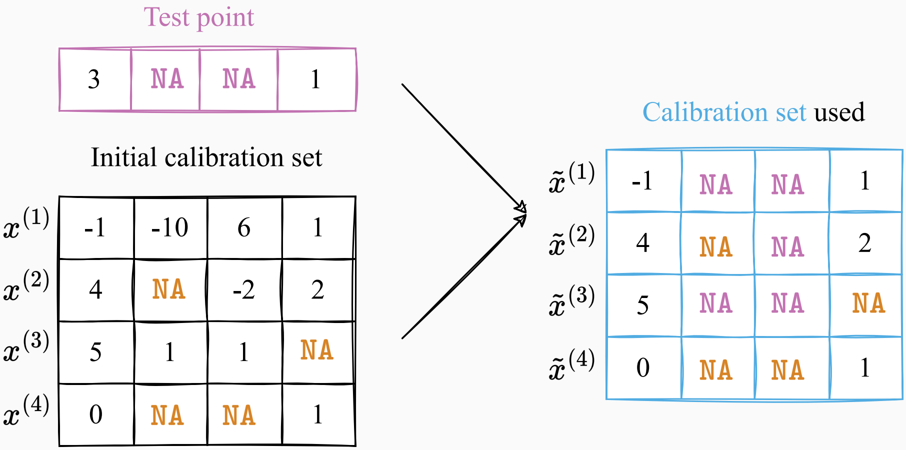


Warning: 2^d possible masks

$\Rightarrow$ Splitting the calibration set by mask *(Mondrian type)* is infeasible (lack of data)!



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



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

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

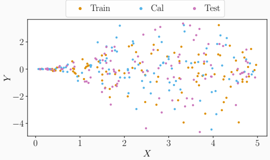
Calibration set used

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

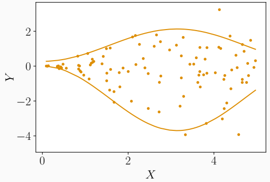
$\#\mathrm{Cal}^{\mathrm{M}^{(\text{test})}}$ observations

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`

CQR-MDA with exact masking in words

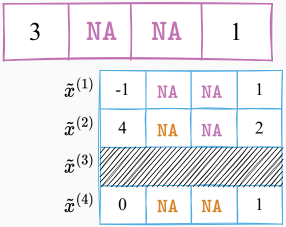
1. Split the training set into a `proper training set` and `calibration set`
2. Train the imputation function on the `proper training set`
3. Impute the `proper training set`
4. Train the quantile regressors on the `imputed proper training set`
5. For a test point $(X^{(n+1)}, M^{(n+1)})$:

3	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. $M^{(k)} \subset M^{(n+1)}$



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`

3	NA	NA	1

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

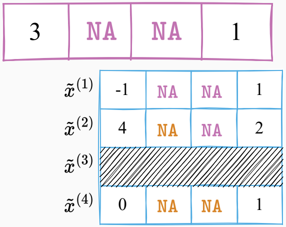
CQR-MDA with exact masking in words

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. $q_{1-\alpha}(\mathcal{S})$



CQR-MDA with exact masking in words

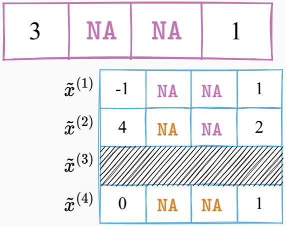
1. Split the training set into a `proper training set` and `calibration set`
2. Train the imputation function on the `proper training set`
3. Impute the `proper training set`
4. Train the quantile regressors on the `imputed proper training set`
5. For a test point $\left(X^{(n+1)}, M^{(n+1)}\right)$:

   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`

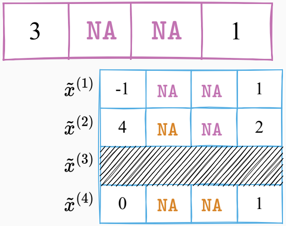
   5.3 Compute the calibration correction, i.e. $q_{1-\alpha}(\mathcal{S})$

   5.4 Impute the test point



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-Exact achieves Mask-Conditional-Validity (MCV)

Theorem (CP-MDA-Exact achieves MCV, Z. et al. (2023a))

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

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

and if additionally the scores are almost surely distinct:

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

What if we kept all observations?

Test point

3	NA	NA	1

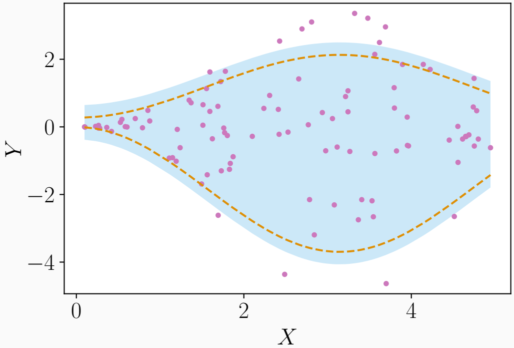
Initial calibration set

$x^{(1)}$	-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

What if we kept all observations?



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

- Build

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

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

$\rightsquigarrow$ similar motivation than Barber et al. (2021)[8] and Gupta et al. (2022)[9].

[8] *Predictive inference with the jackknife+*, The Annals of Statistics

[9] *Nested conformal prediction and quantile out-of-bag ensemble methods*, Pattern Recognition

CQR-MDA with nested masking in words

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

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

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

CQR-MDA with nested masking in words

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

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

   5.2 Impute the new `calibration set`

   5.3 For each augmented calibration point k:

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

CQR-MDA with nested masking in words

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

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

   5.2 Impute the new `calibration set`

   5.3 For each augmented calibration point k:

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

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

CQR-MDA with nested masking in words

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

3	NA	NA	1

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

5.2 Impute the new `calibration set`

5.3 For each augmented calibration point k:

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

5.3.2 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

CQR-MDA with nested masking in words

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

3	NA	NA	1

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

  5.2 Impute the new `calibration set`

  5.3 For each augmented calibration point k:

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

    5.3.2 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

    5.3.3 Compute the corrected prediction interval:

$$[\widehat{QR}_{\alpha/2}(\tilde{X}^{(n+1),k}) - S^{(k)}; \widehat{QR}_{1-\alpha/2}(\tilde{X}^{(n+1),k}) + S^{(k)}] := \left[Z^{(k)}_{\text{lower}}; Z^{(k)}_{\text{upper}}\right]$$

CQR-MDA with nested masking in words

5. 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

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

5.2 Impute the new `calibration set`

5.3 For each augmented calibration point k:

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

5.3.2 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

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

5.4 Compute the quantiles $q_\alpha(\{Z^{(k)}_{\text{lower}}\}_{k\in\mathrm{Cal}})$ and $q_{1-\alpha}(\{Z^{(k)}_{\text{upper}}\}_{k\in\mathrm{Cal}})$

CQR-MDA with nested masking in words

5. 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

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

5.2 Impute the new `calibration set`

5.3 For each augmented calibration point k:

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

5.3.2 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

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

5.4 Compute the quantiles $q_\alpha(\{Z^{(k)}_{\text{lower}}\}_{k\in\mathrm{Cal}})$ and $q_{1-\alpha}(\{Z^{(k)}_{\text{upper}}\}_{k\in\mathrm{Cal}})$

5.5 Predict $[q_\alpha(\{Z^{(k)}_{\text{lower}}\}_{k\in\mathrm{Cal}}); q_{1-\alpha}(\{Z^{(k)}_{\text{upper}}\}_{k\in\mathrm{Cal}})]$

MDA-Nested is Marginally Valid (MV)

Theorem (CP-MDA-Nested marginal validity, Z. et al. (2023b))

If the data is exchangeable, then for almost all imputation function CP-MDA-Nested is such that:

$$\mathbb{P}\left(Y \in \widehat{C}_\alpha\left(X, M\right)\right) \geq 1 - 2\alpha.$$

MDA-Nested is Marginally Valid (MV)

Theorem (CP-MDA-Nested marginal validity, Z. et al. (2023b))

If the data is exchangeable, then for almost all imputation function CP-MDA-Nested is such that:

$$\mathbb{P}\left(Y \in \widehat{C}_\alpha\left(X, M\right)\right) \geq 1 - 2\alpha.$$

- ✓ Any missing mechanism (no need to assume $M \perp\!\!\!\perp X$)
- ✓ Does not require $(Y \perp\!\!\!\perp M)\,|X$
- ✗ Marginal guarantee

MDA-Nested is Marginally Valid (MV)

Theorem (CP-MDA-Nested marginal validity, Z. et al. (2023b))

If the data is exchangeable, then for almost all imputation function CP-MDA-Nested is such that:

$$\mathbb{P}\left(Y \in \widehat{C}_\alpha\left(X, M\right)\right) \geq 1 - 2\alpha.$$

- ✓ Any missing mechanism (no need to assume $M \perp\!\!\!\perp X$)
- ✓ Does not require $(Y \perp\!\!\!\perp M) | X$
- ✗ Marginal guarantee

Proof element: based on Jackknife+ ideas (Barber et al., 2021).

Leaving-out the k-th data point to predict on the l-th data point

$\leftrightarrow$

Apply the mask of the k-th data point to the l-th data point on which you predict

MDA-Nested (nearly) achieves Mask-Conditional-Validity (MCV)

Stochastic domination of the quantiles (SDQ)

Let $(\mathring{m}, \breve{m}) \in \left(\{0,1\}^d\right)^2$. If $\mathring{m} \subset \breve{m}$ then for any $\delta \in [0, 0.5]$: $q_{1-\delta/2}^{Y|(X_{\text{obs}(\mathring{m})}, M=\mathring{m})} \leq q_{1-\delta/2}^{Y|(X_{\text{obs}(\breve{m})}, M=\breve{m})}$, and $q_{\delta/2}^{Y|(X_{\text{obs}(\mathring{m})}, M=\mathring{m})} \geq q_{\delta/2}^{Y|(X_{\text{obs}(\breve{m})}, M=\breve{m})}$.

$\rightsquigarrow$ predictive uncertainty increases with bigger masks.

MDA-Nested (nearly) achieves Mask-Conditional-Validity (MCV)

Stochastic domination of the quantiles (SDQ)

Let $(\mathring{m}, \breve{m}) \in \left(\{0,1\}^d\right)^2$. If $\mathring{m} \subset \breve{m}$ then for any $\delta \in [0, 0.5]$: $q_{1-\delta/2}^{Y|(X_{\text{obs}(\mathring{m})}, M=\mathring{m})} \leq q_{1-\delta/2}^{Y|(X_{\text{obs}(\breve{m})}, M=\breve{m})}$, and $q_{\delta/2}^{Y|(X_{\text{obs}(\mathring{m})}, M=\mathring{m})} \geq q_{\delta/2}^{Y|(X_{\text{obs}(\breve{m})}, M=\breve{m})}$.

$\rightsquigarrow$ predictive uncertainty increases with bigger masks.

Theorem (CP-MDA-Nested (nearly) achieves MCV, Z. et al. (2023a))

If i) the data is exchangeable, ii) $M \perp\!\!\!\perp X$, iii) $(Y \perp\!\!\!\perp M)|X$, iv) SDQ holds, then for almost all imputation function "CP-MDA-Nested" is s.t. for any $m \in \{0,1\}^d$:

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

MDA-Nested (nearly) achieves Mask-Conditional-Validity (MCV)

Stochastic domination of the quantiles (SDQ)

Let $(\mathring{m}, \breve{m}) \in (\{0,1\}^d)^2$. If $\mathring{m} \subset \breve{m}$ then for any $\delta \in [0, 0.5]$: $q_{1-\delta/2}^{Y|(X_{\text{obs}(\mathring{m})}, M=\mathring{m})} \leq q_{1-\delta/2}^{Y|(X_{\text{obs}(\breve{m})}, M=\breve{m})}$, and $q_{\delta/2}^{Y|(X_{\text{obs}(\mathring{m})}, M=\mathring{m})} \geq q_{\delta/2}^{Y|(X_{\text{obs}(\breve{m})}, M=\breve{m})}$.

$\rightsquigarrow$ predictive uncertainty increases with bigger masks.

Theorem (CP-MDA-Nested (nearly) achieves MCV, Z. et al. (2023a))

If i) the data is exchangeable, ii) $M \perp\!\!\!\perp X$, iii) $(Y \perp\!\!\!\perp M)|X$, iv) SDQ holds, then for almost all imputation function "CP-MDA-Nested" is s.t. for any $m \in \{0,1\}^d$:

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

Change on MDA-Nested: outputs any $[q_\alpha(\{Z_{\text{lower}}^{(k)}\}_{k \in \mathrm{Cal}^{\breve{m}}}); q_{1-\alpha}(\{Z_{\text{upper}}^{(k)}\}_{k \in \mathrm{Cal}^{\breve{m}}})]$, where $\breve{m}$ is randomly[10] selected such that $m \subset \breve{m}$.

[10] The randomness may depend on $\#\mathrm{Cal}^{\breve{m}}$.

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

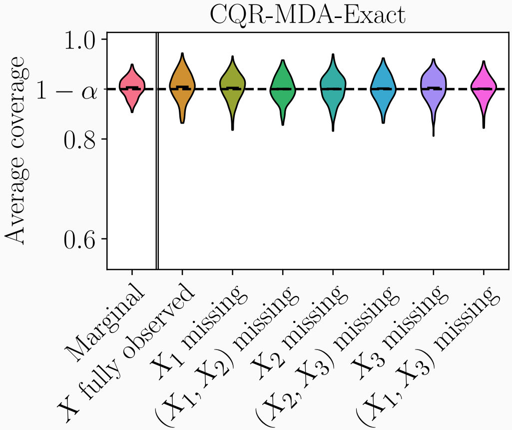
temporary test points

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

MDA achieves Mask-Conditional-Validity (MCV)

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

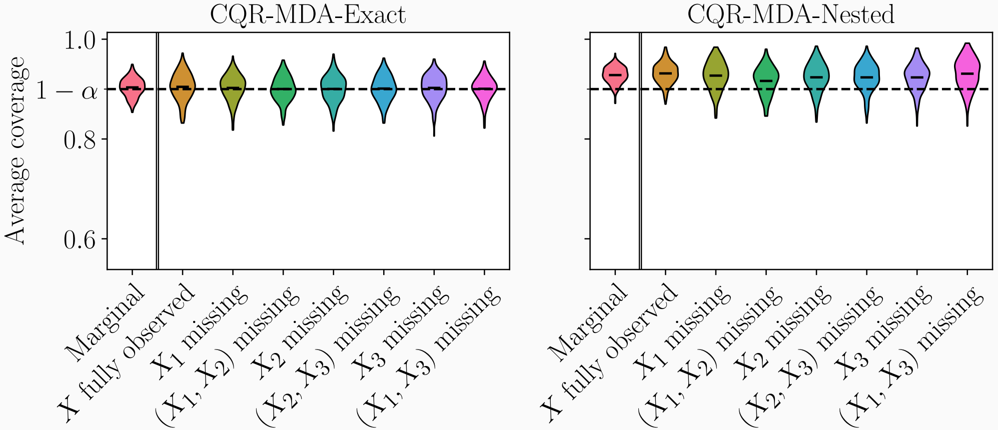
$\beta = (1, 2, -1)^T$, $\varepsilon \perp\!\!\!\perp X$, X and ε Gaussian, 20% uniform MCAR missing values.



MDA achieves Mask-Conditional-Validity (MCV)

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

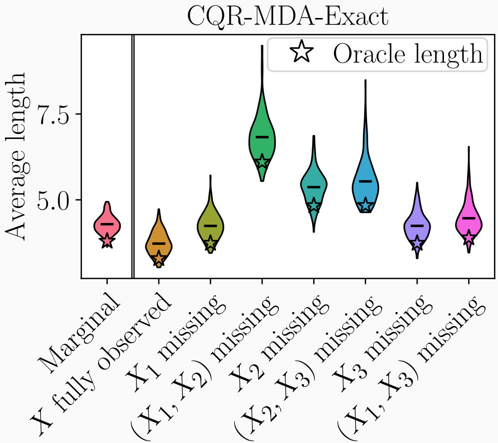
$\beta = (1, 2, -1)^T$, $\varepsilon \perp\!\!\!\perp X$, X and ε Gaussian, 20% uniform MCAR missing values.



MDA achieves (MCV) in an informative way

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

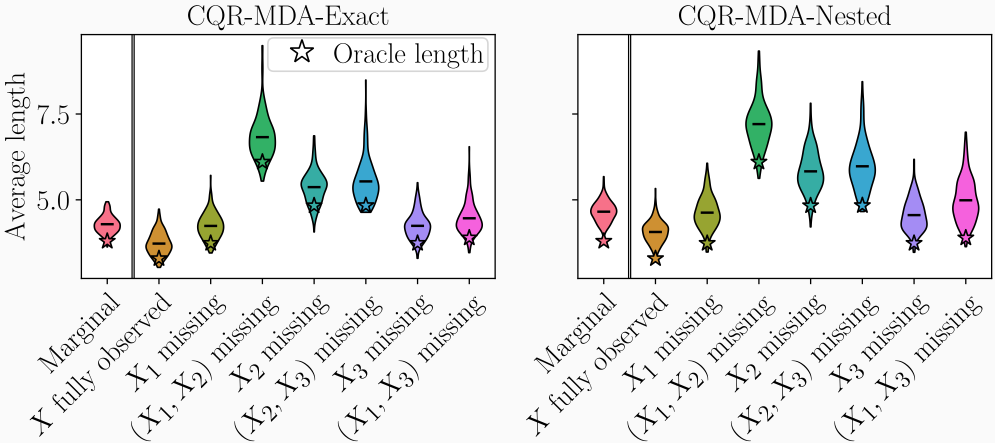
$\beta = (1, 2, -1)^T$, $\varepsilon \perp\!\!\!\perp X$, X and ε Gaussian, 20% uniform MCAR missing values.



MDA achieves (MCV) in an informative way

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

$\beta = (1, 2, -1)^T$, $\varepsilon \perp\!\!\!\perp X$, X and ε Gaussian, 20% uniform MCAR missing values.



Some settings

- Imputation by iterative ridge ($\sim$ conditional expectation)

Some settings

- Imputation by iterative ridge ($\sim$ conditional expectation)
- **Concatenate the mask in the features**

Some settings

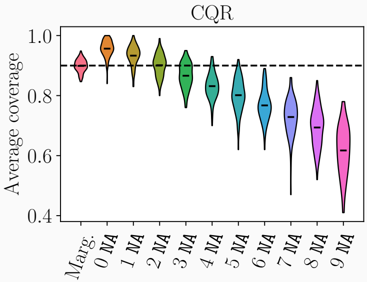
- Imputation by iterative ridge ($\sim$ conditional expectation)
- **Concatenate the mask in the features**
- Neural network, fitted to minimize the pinball loss

- Imputation by iterative ridge ($\sim$ conditional expectation)
- **Concatenate the mask in the features**
- Neural network, fitted to minimize the pinball loss
- (Semi)-synthetic experiments:

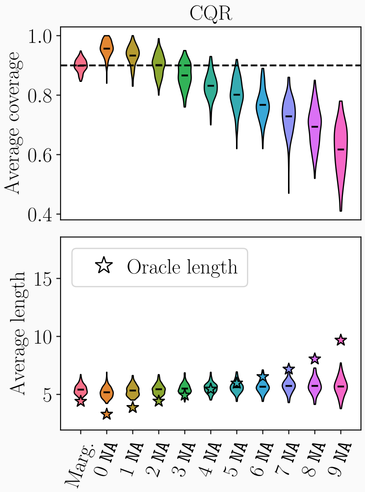
Some settings

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

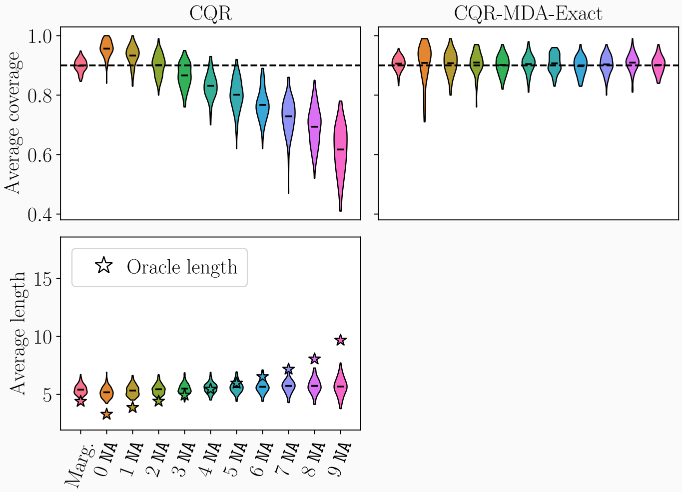
Synthetic experiments (Gaussian linear model, $d = 10$)



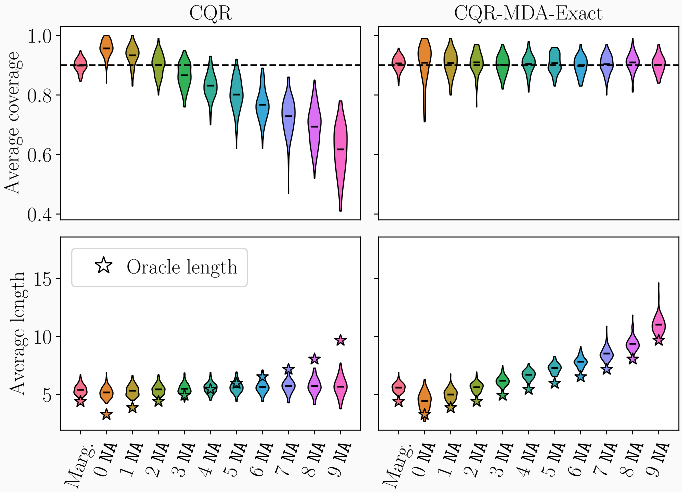
Synthetic experiments (Gaussian linear model, $d = 10$)



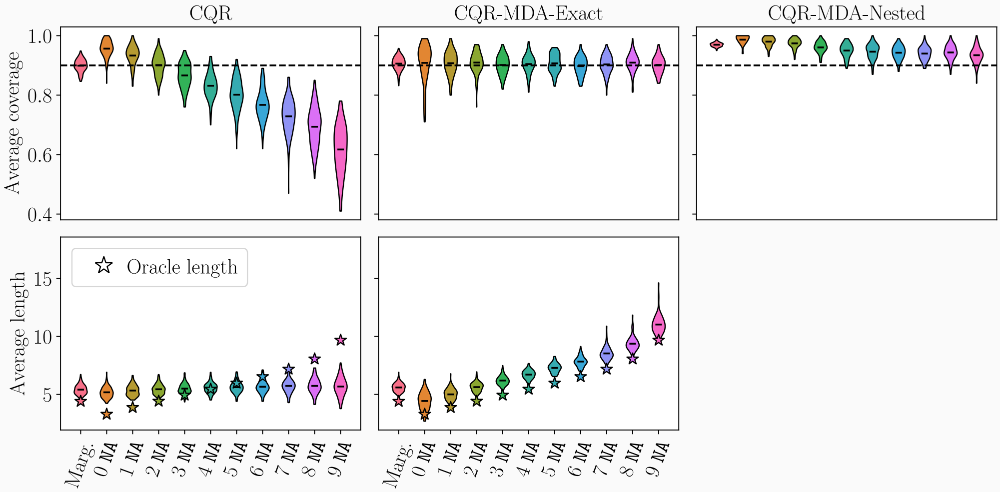
Synthetic experiments (Gaussian linear model, $d = 10$)



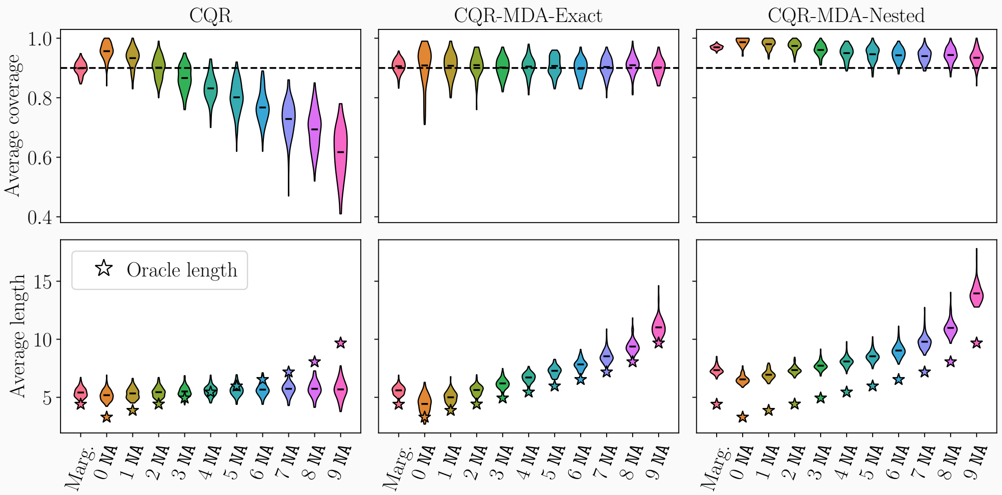
Synthetic experiments (Gaussian linear model, $d = 10$)



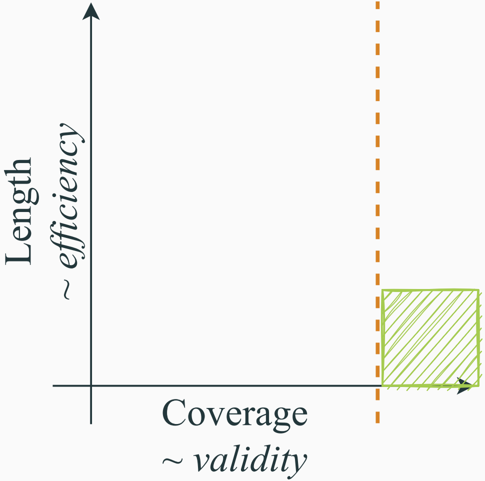
Synthetic experiments (Gaussian linear model, $d = 10$)

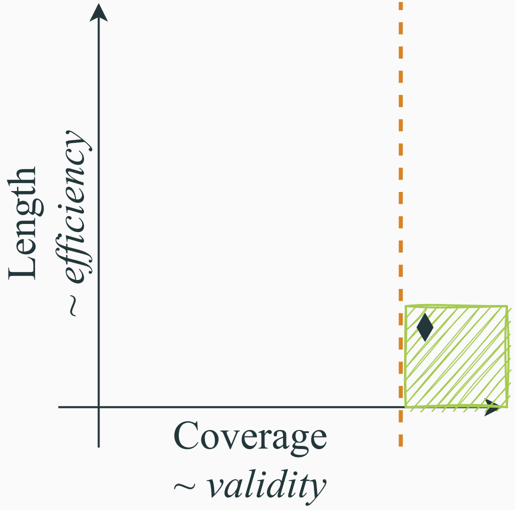


Synthetic experiments (Gaussian linear model, $d = 10$)



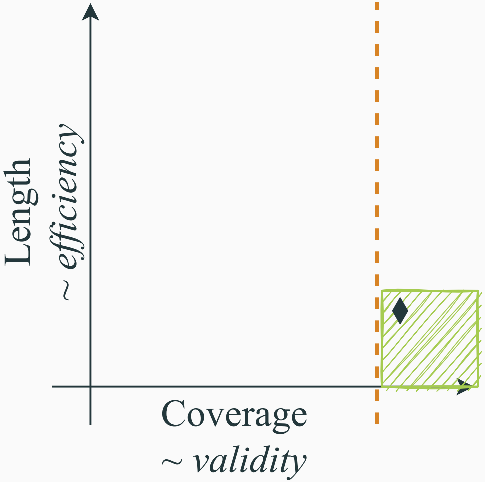
Before more experiments, visualisation





♦ : marginal coverage, i.e. $\mathbb{P}(Y \in \hat{C}_\alpha(X, M))$

Before more experiments, visualisation



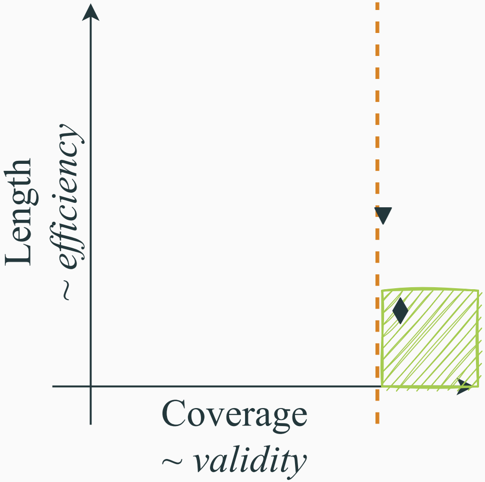
- ♦ : marginal coverage, i.e.

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

- ▼ : lowest coverage, i.e.

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

Before more experiments, visualisation

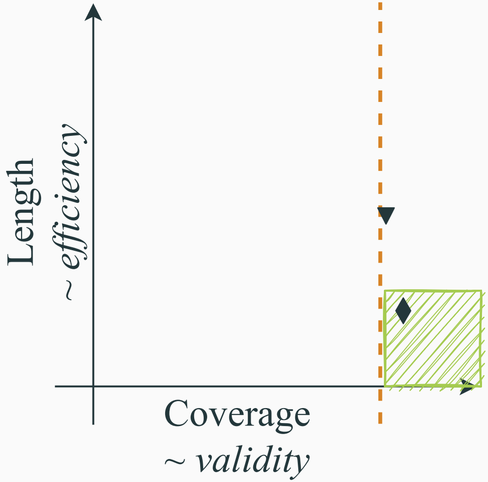


♦ : marginal coverage, i.e.

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

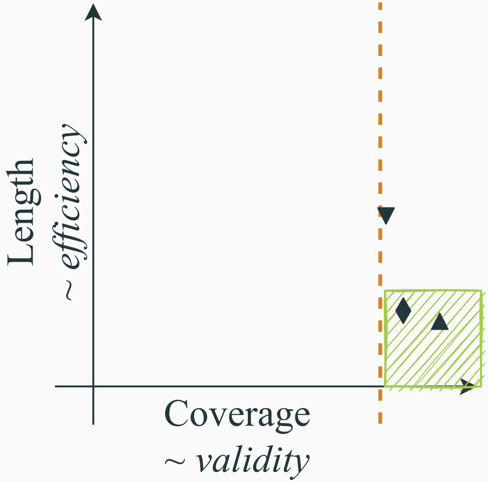
▼ : lowest coverage, i.e.

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



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

Before more experiments, visualisation



♦ : marginal coverage, i.e.

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

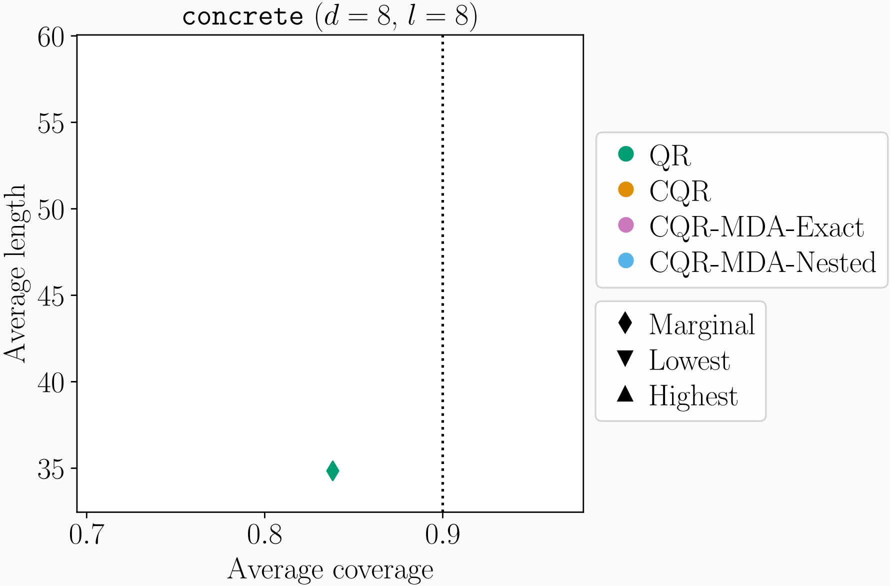
▼ : lowest coverage, i.e.

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

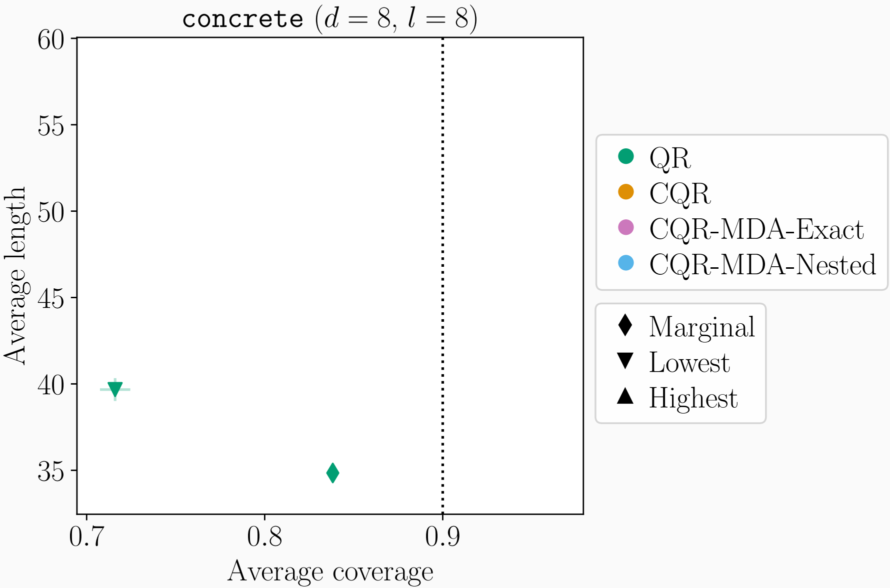
▲ : highest coverage, i.e.

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

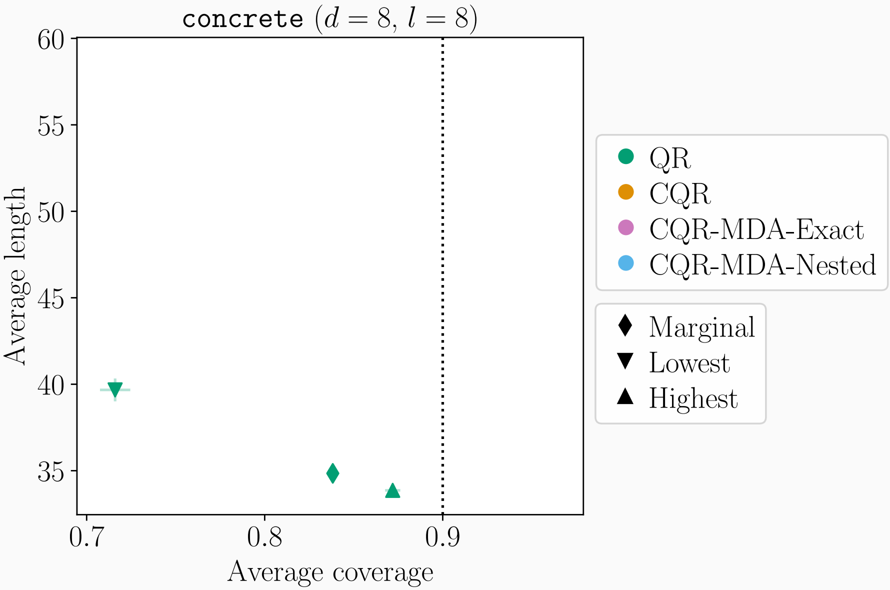
Semi-synthetic experiments



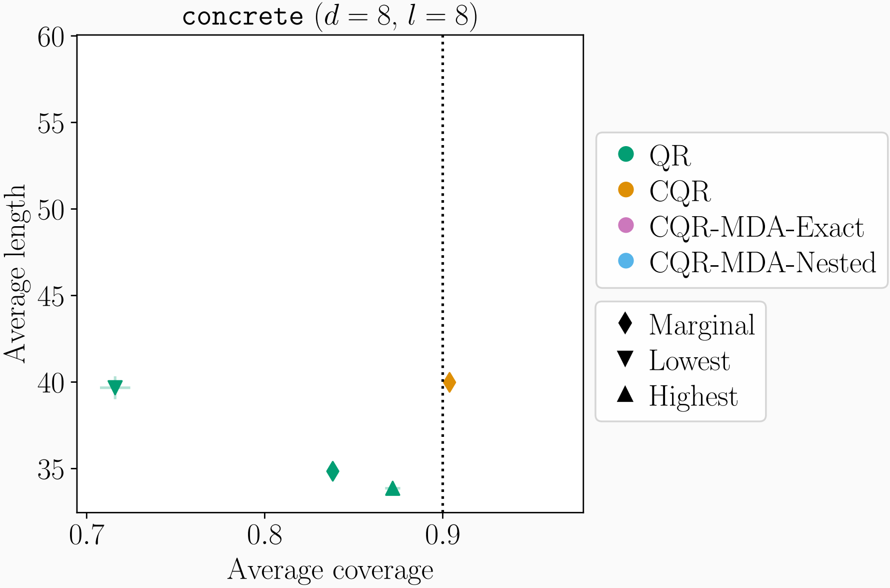
Semi-synthetic experiments



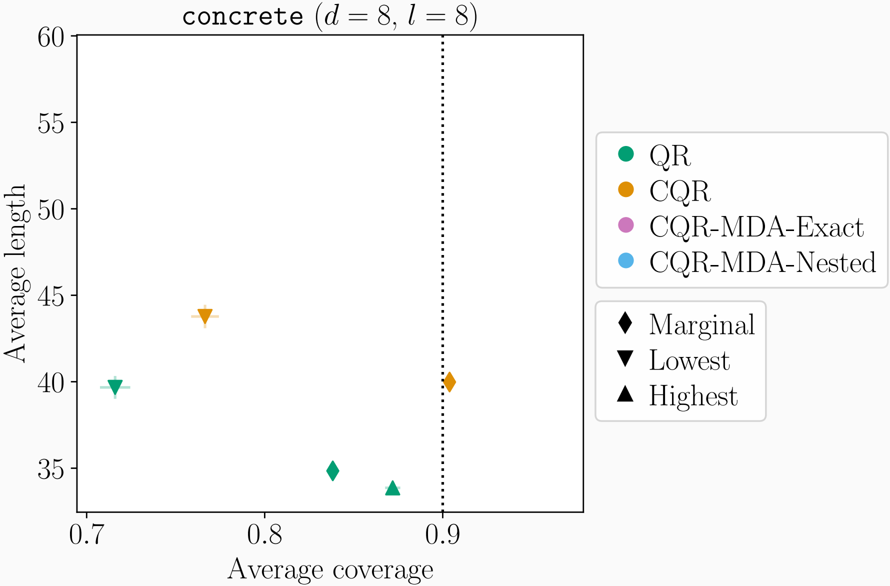
Semi-synthetic experiments



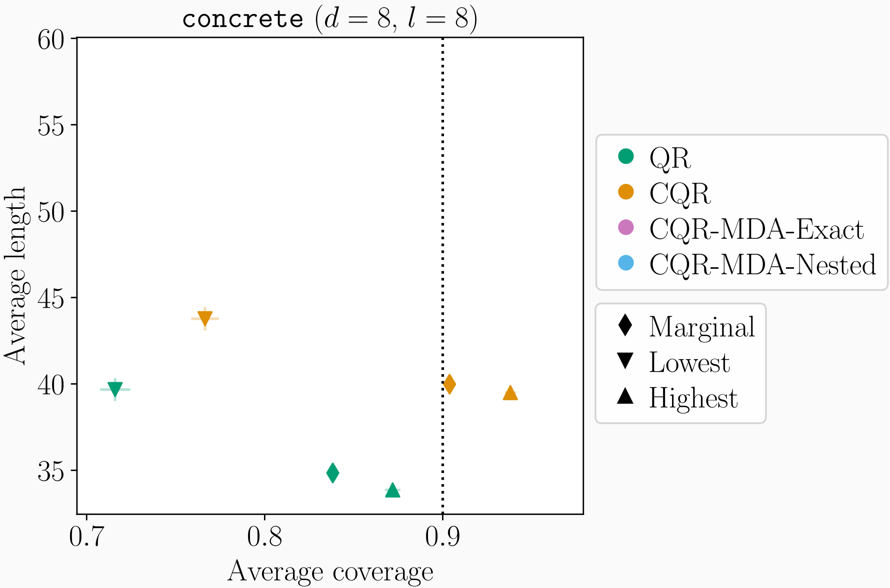
Semi-synthetic experiments



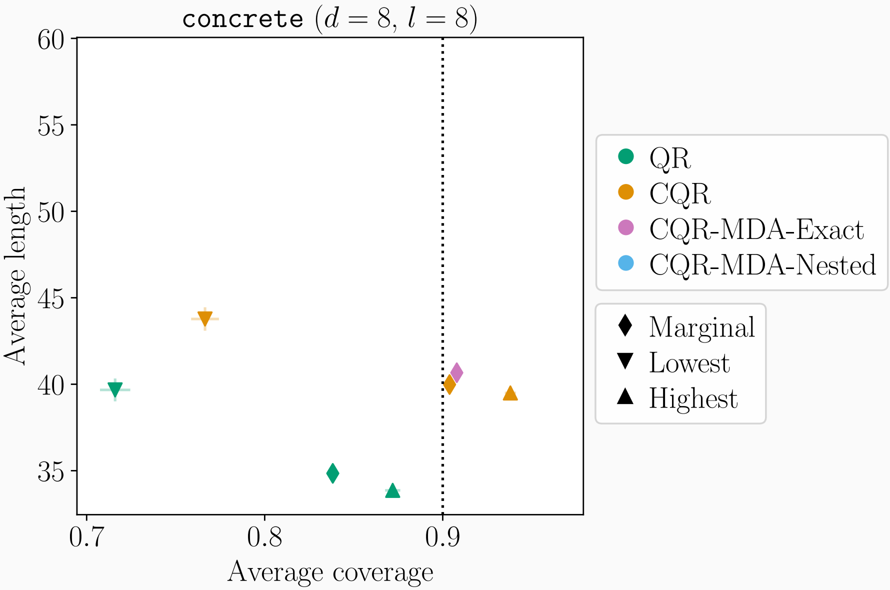
Semi-synthetic experiments

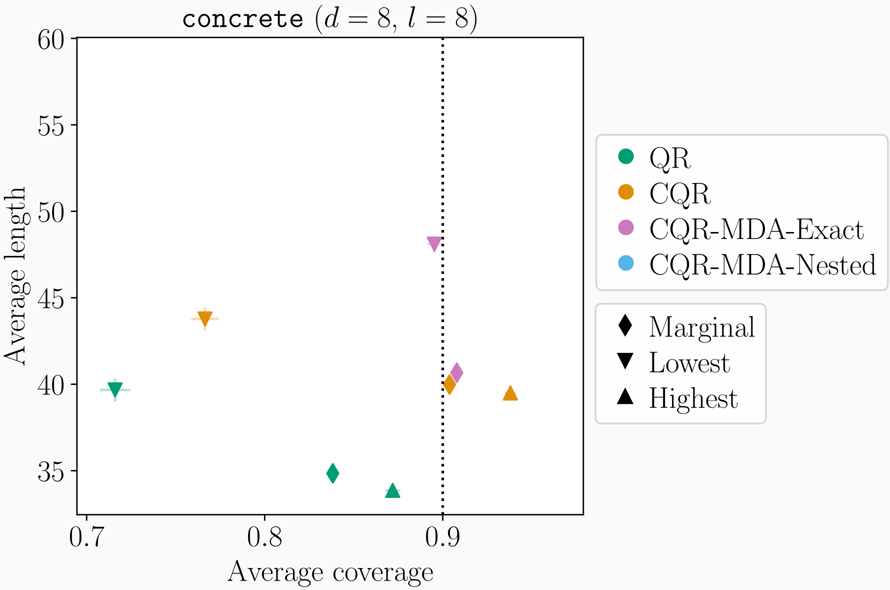


Semi-synthetic experiments

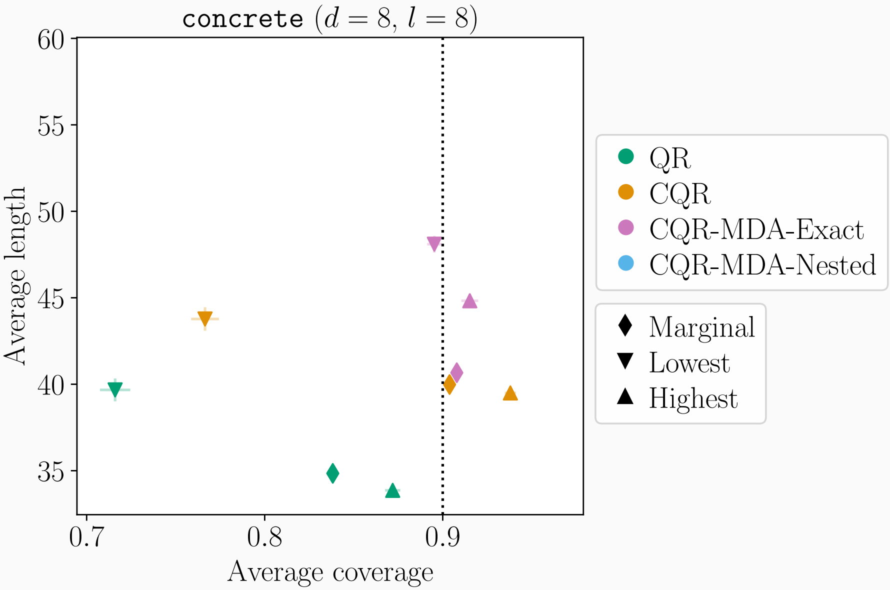


Semi-synthetic experiments

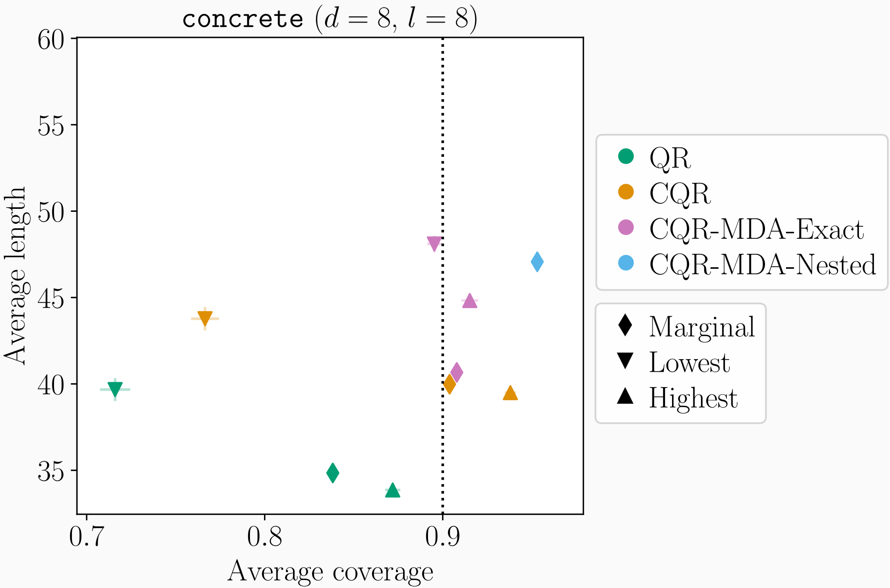



concrete ($d = 8$, $l = 8$)
60
55
50
45
40
35
Average length
0.7
0.8
0.9
Average coverage
QR
CQR
CQR-MDA-Exact
CQR-MDA-Nested
Marginal
Lowest
Highest

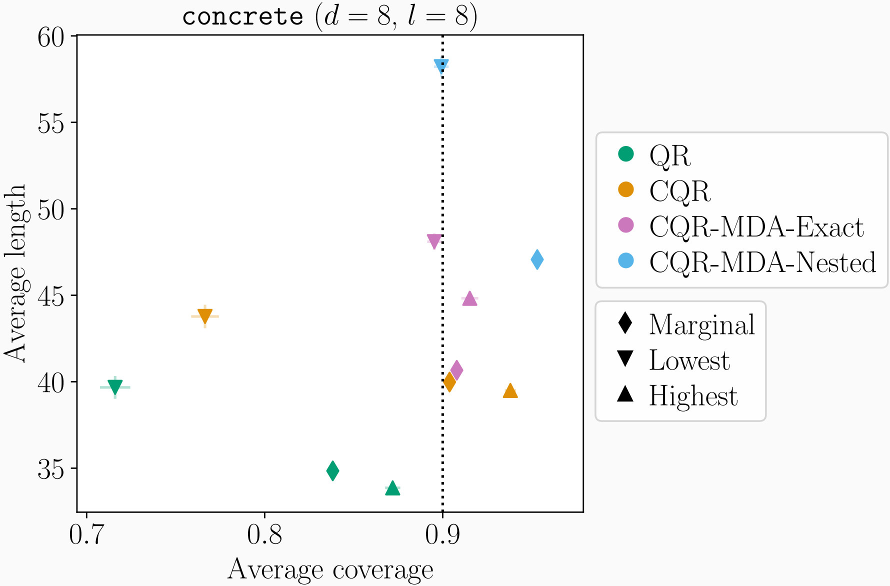
Semi-synthetic experiments



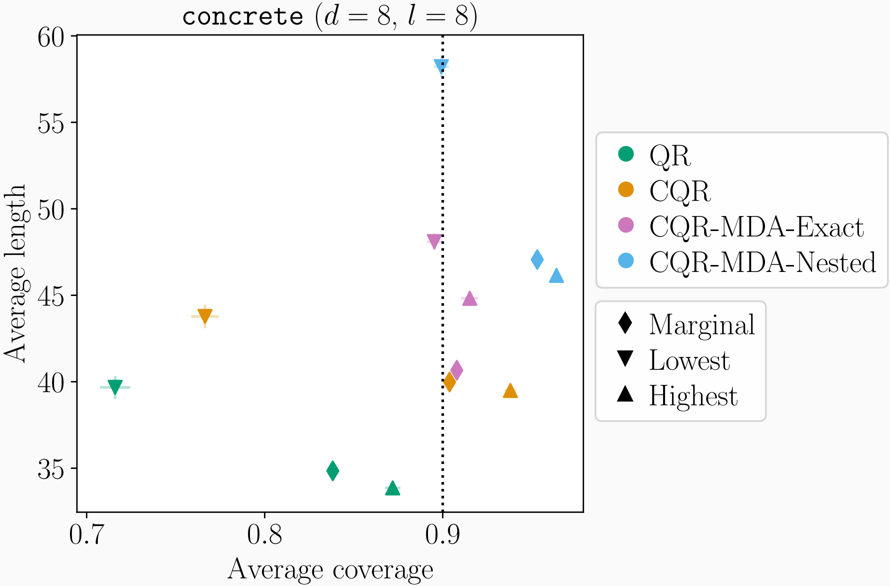
Semi-synthetic experiments

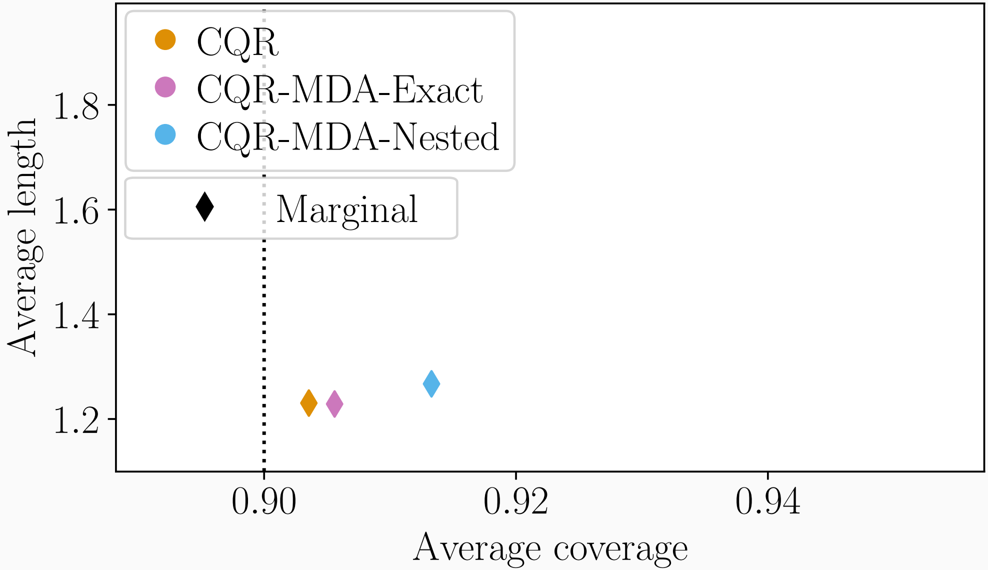


Semi-synthetic experiments



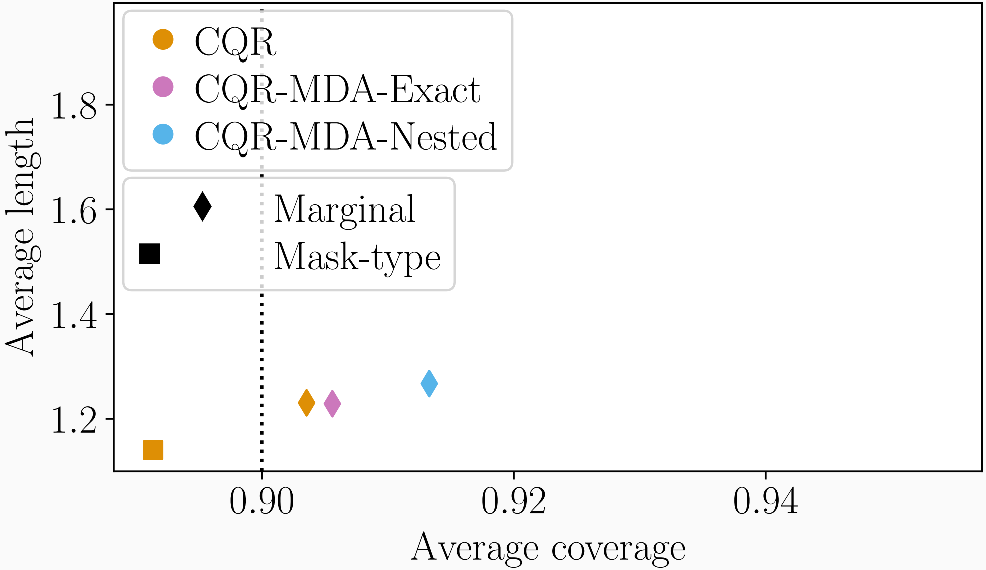
Semi-synthetic experiments

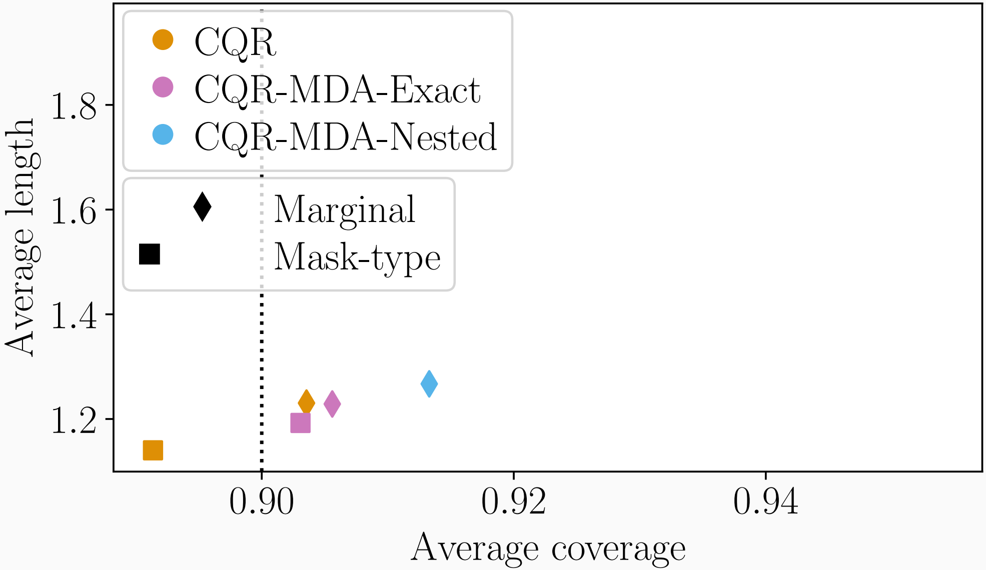


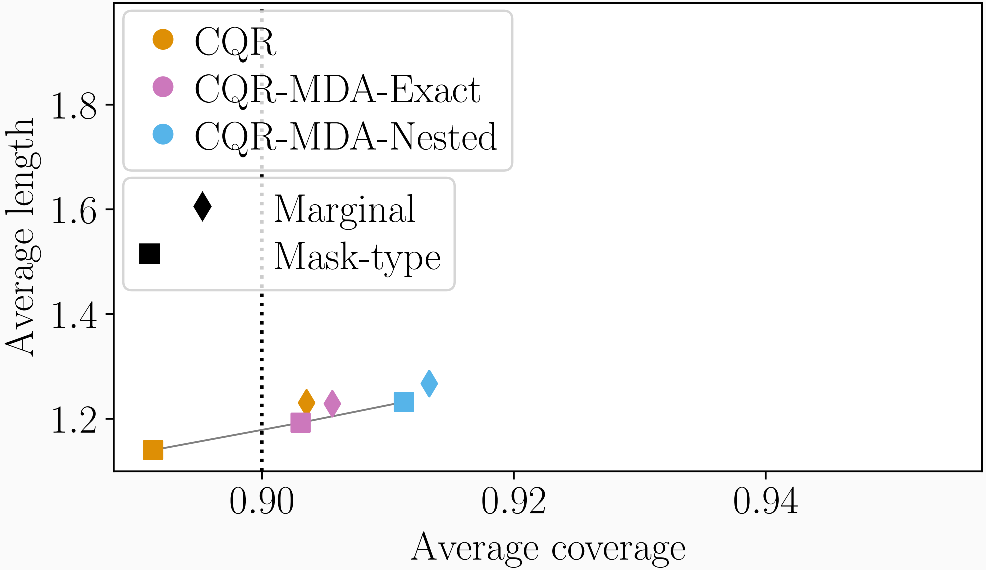


CQR
CQR-MDA-Exact
CQR-MDA-Nested
Marginal
Average length
1.8
1.6
1.4
1.2
0.90
0.92
0.94
Average coverage

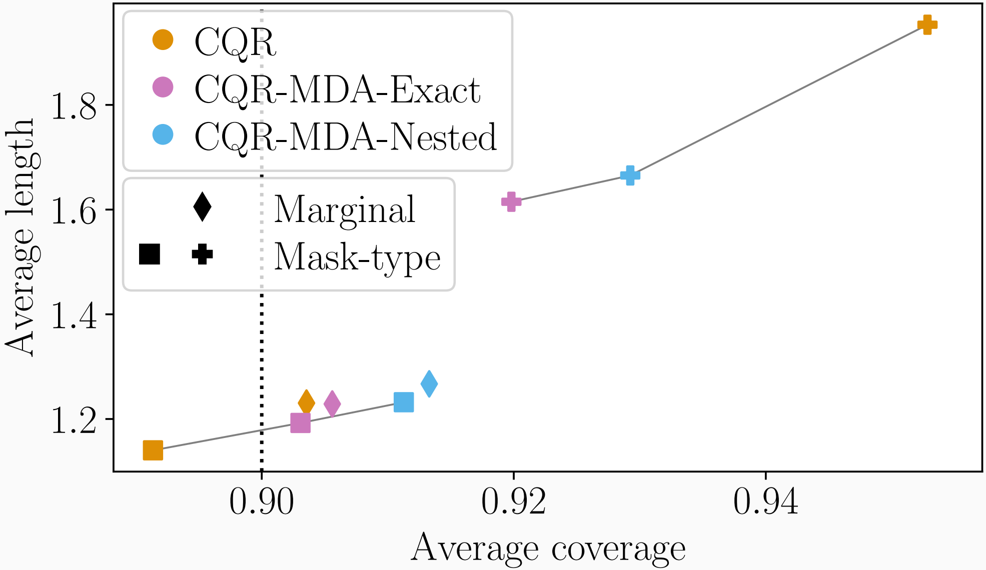
Real data experiment: TraumaBase®, critical care medicine



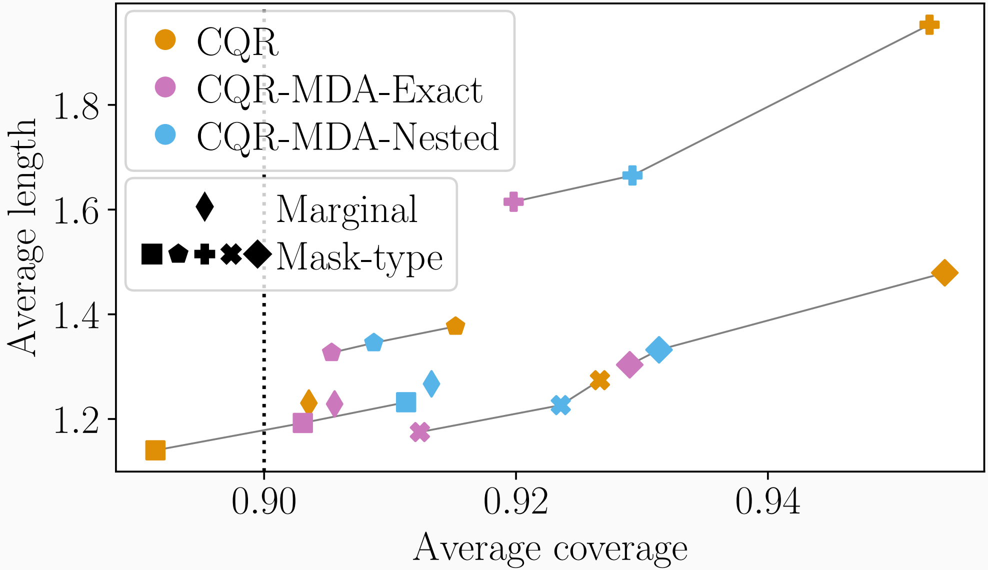

CQR
CQR-MDA-Exact
CQR-MDA-Nested
Marginal
Mask-type
Average length
1.2
1.4
1.6
1.8
0.90
0.92
0.94
Average coverage


CQR
CQR-MDA-Exact
CQR-MDA-Nested
Marginal
Mask-type
Average length
1.8
1.6
1.4
1.2
0.90
0.92
0.94
Average coverage

Real data experiment: TraumaBase®, critical care medicine



Real data experiment: TraumaBase®, critical care medicine



Introduction to missing values

Quantifying predictive uncertainty with missing values

Conclusion

Take-home-messages

- CP marginal guarantees hold on the imputed data set.

Take-home-messages

- CP marginal guarantees hold on the imputed data set.
- Missingness introduces additional heteroskedasticity, creating a need for quantile regression based methods.

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.

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**.

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).

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).
- Extension: consistency of universal quantile learner when chained with almost any imputation function.

Perspectives/connection to other works

- Investigate alternative methods relying on trade-offs between MDA-Exact and MDA-Nested

Perspectives/connection to other works

- Investigate alternative methods relying on trade-offs between MDA-Exact and MDA-Nested
- Relationship with Gibbs et al. (2023)[11]
  - ✓ Beyond MCAR
  - ✗ Upper bound in $\frac{2^d}{(n+1)\mathbb{P}_M(m)}$: high value for less probable masks
  - ↪ MCV are non-overlapping groups: boils down to splitting the calibration set!

[11] *Conformal Prediction With Conditional Guarantees*

Perspectives/connection to other works

- Investigate alternative methods relying on trade-offs between MDA-Exact and MDA-Nested
- Relationship with Gibbs et al. (2023)[11]
  - ✓ Beyond MCAR
  - ✗ Upper bound in $\frac{2^d}{(n+1)\mathbb{P}_M(m)}$: high value for less probable masks
  - ↪ MCV are non-overlapping groups: boils down to splitting the calibration set!

[11] *Conformal Prediction With Conditional Guarantees*

Perspectives/connection to other works

- Investigate alternative methods relying on trade-offs between MDA-Exact and MDA-Nested
- Relationship with Gibbs et al. (2023)[11]
  - ✓ Beyond MCAR
  - ✗ Upper bound in $\frac{2^d}{(n+1)\mathbb{P}_M(m)}$: high value for less probable masks
  - ↪ MCV are non-overlapping groups: boils down to splitting the calibration set!
- Quantify the impact of the imputation's choice on Quantile Regression quality in finite sample

[11] *Conformal Prediction With Conditional Guarantees*

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. *ICML*.

Barber, R. F., Candès, E. J., Ramdas, A., and Tibshirani, R. J. (2021). Predictive inference with the jackknife+. *The Annals of Statistics*, 49(1).

Gibbs, I., Cherian, J. J., and Candès, E. J. (2023). Conformal prediction with conditional guarantees.

Gupta, C., Kuchibhotla, A. K., and Ramdas, A. (2022). Nested conformal prediction and quantile out-of-bag ensemble methods. *Pattern Recognition*, 127.

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*.

Romano, Y., Patterson, E., and Candès, E. (2019). Conformalized Quantile Regression. *NeurIPS*.

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

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

Z. , M., Dieuleveut, A., Josse, J., and Romano, Y. (2023a). Conformal prediction with missing values. *ICML*.

Z. , M., Dieuleveut, A., Josse, J., and Romano, Y. (2023b). Predictive uncertainty quantification with missing values. To be submitted.

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

Appendix

Towards asymptotic individualized coverage

Consistency of a universal quantile learner after imputation

Let Φ be an imputation function chosen by the user.

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

Consistency of a universal quantile learner after imputation

Let Φ be an imputation function chosen by the user.

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

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

Consistency of a universal quantile learner after imputation

Let Φ be an imputation function chosen by the user.

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

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

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.

Consistency of a universal quantile learner after imputation

Let Φ be an imputation function chosen by the user.

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

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

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

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

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

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

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

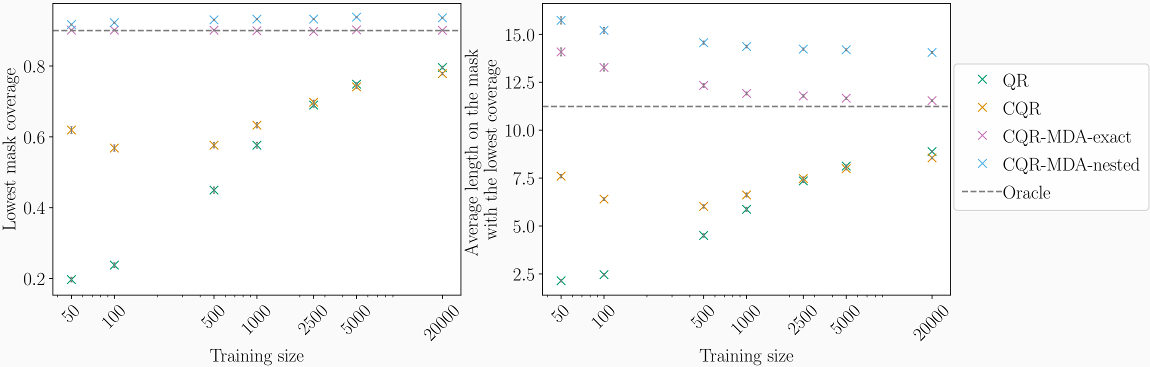
$$(X_1, \cdots, X_{10}) \sim \mathcal{N}\left(\begin{pmatrix} 1 \\ \vdots \\ \vdots \\ 1 \end{pmatrix}, \begin{pmatrix} 1 & 0.8 & \cdots & 0.8 \\ 0.8 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0.8 \\ 0.8 & \cdots & 0.8 & 1 \end{pmatrix}\right).$$

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

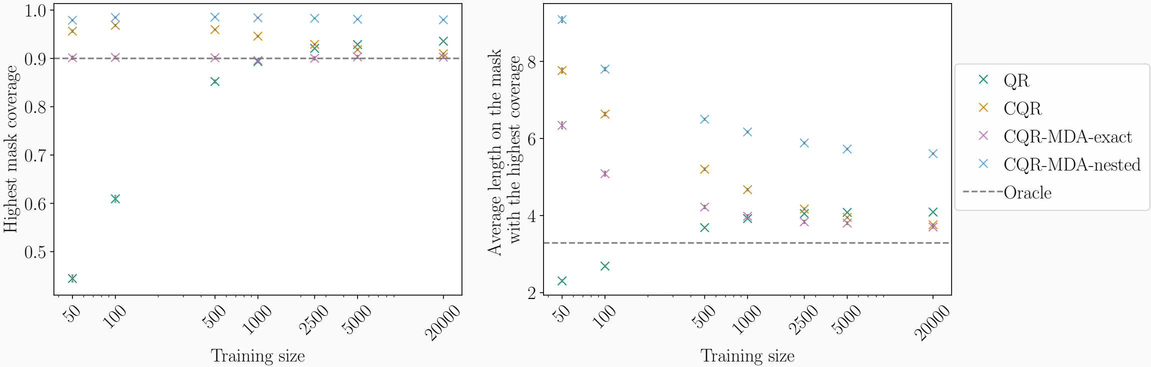
Simulation settings: varying training size

- 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

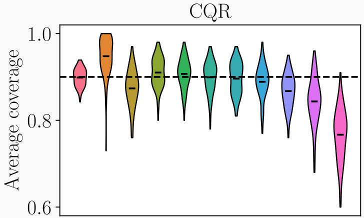
Results on the worst group



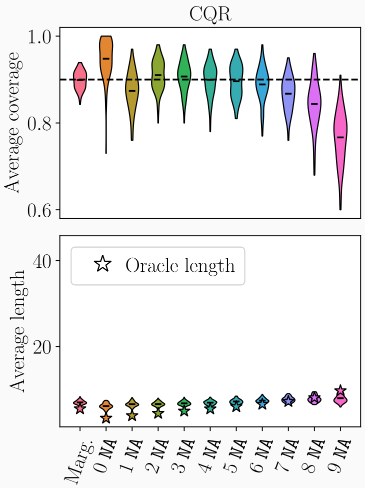
Results on the best group



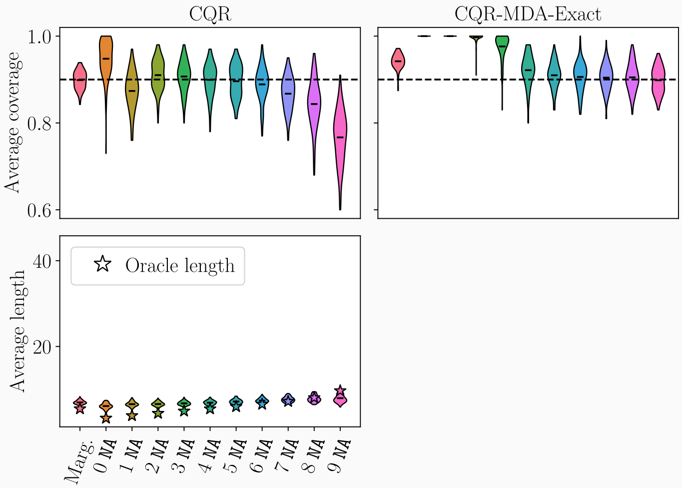
Synthetic experiments, 40% of missing values (Gaussian linear model, $d = 10$)



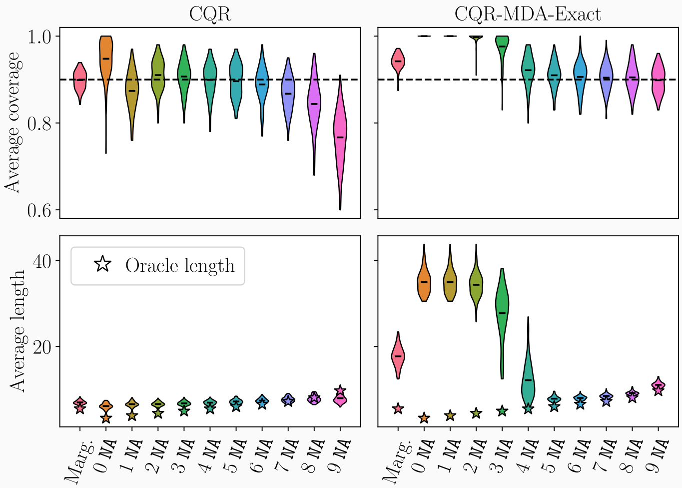
Synthetic experiments, 40% of missing values (Gaussian linear model, $d = 10$)



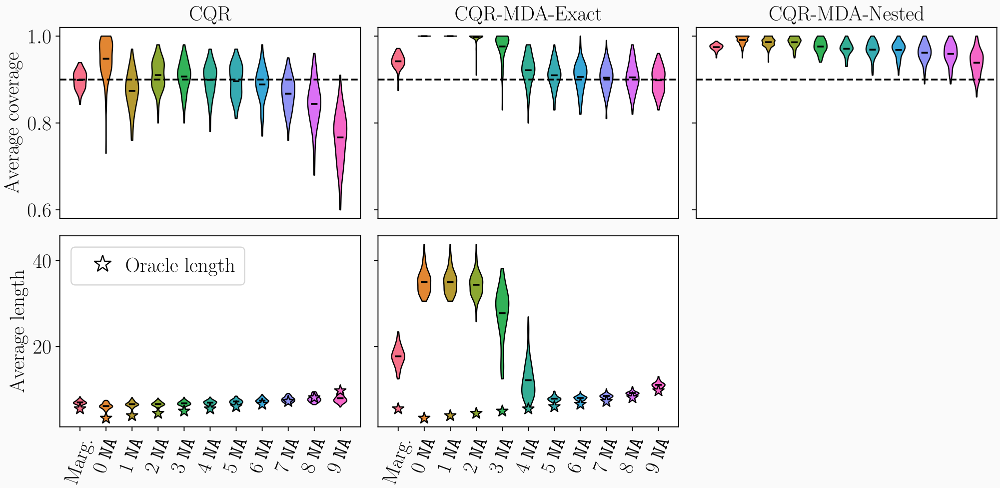
Synthetic experiments, 40% of missing values (Gaussian linear model, $d = 10$)



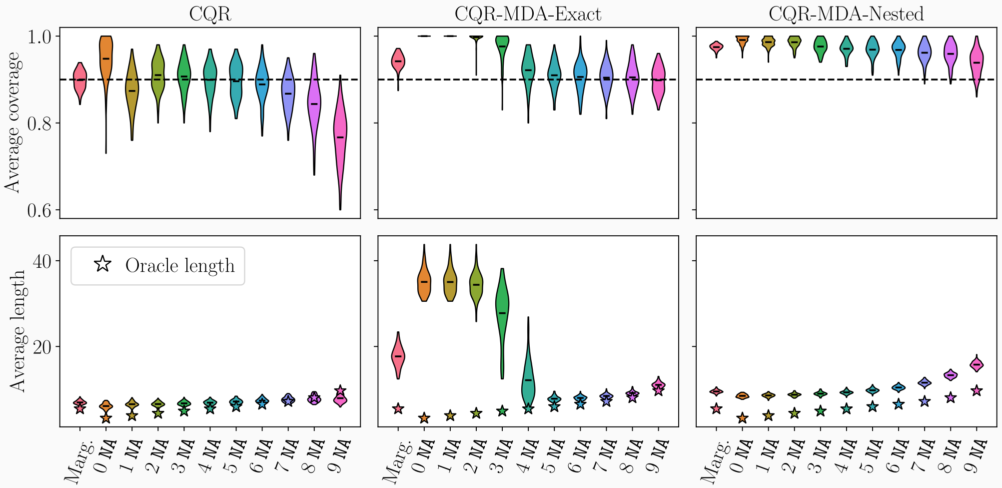
Synthetic experiments, 40% of missing values (Gaussian linear model, $d = 10$)



Synthetic experiments, 40% of missing values (Gaussian linear model, $d = 10$)



Synthetic experiments, 40% of missing values (Gaussian linear model, $d = 10$)

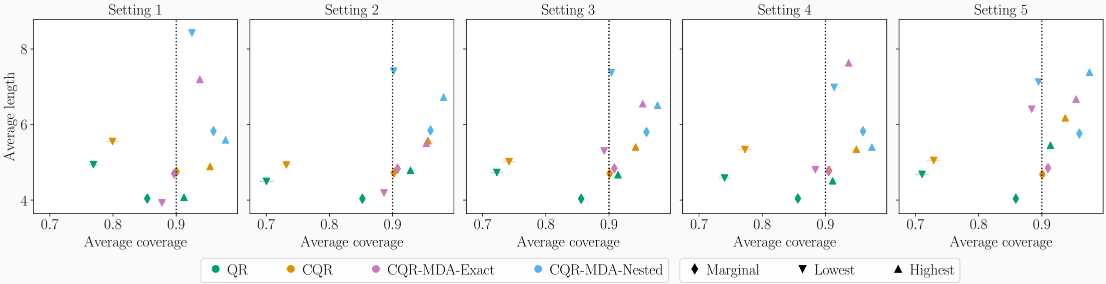


Simulation settings: beyond MCAR

- 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%.

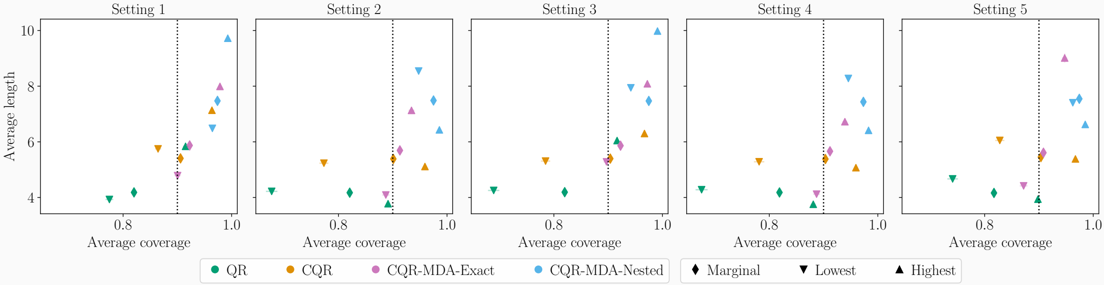
MAR missingness

- 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.



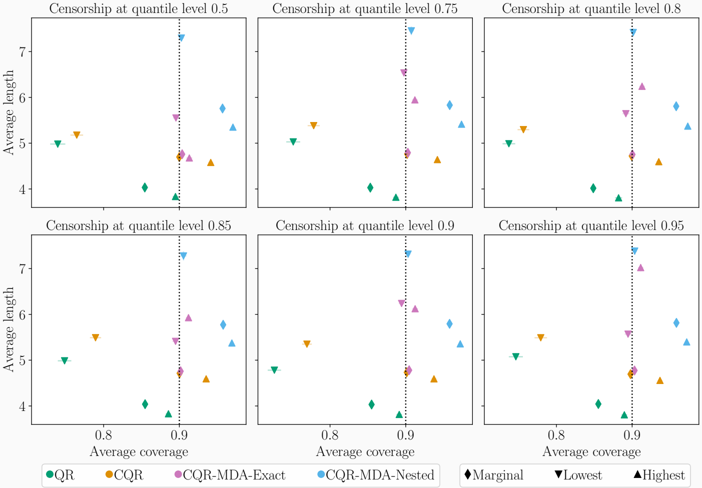
MNAR self masked missingness

- 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.



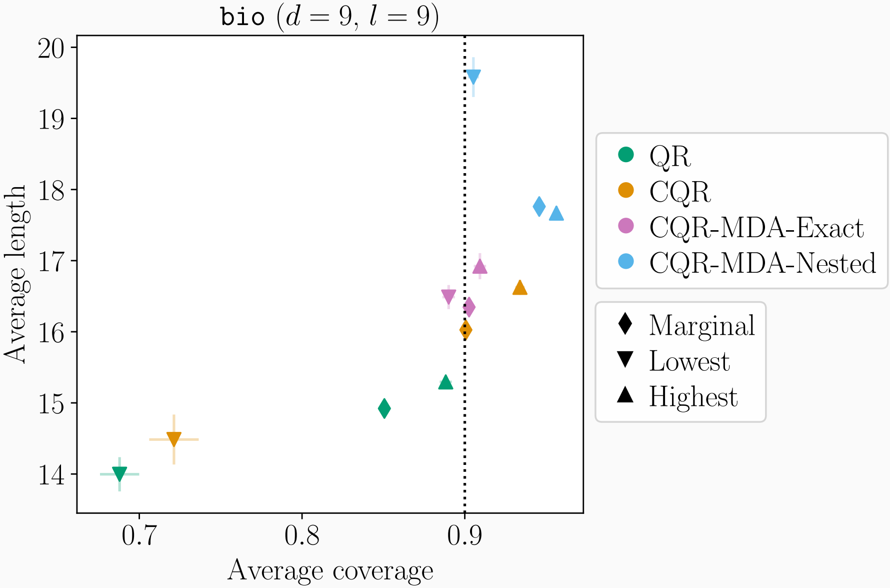
MNAR quantile censorship missingness

- 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.



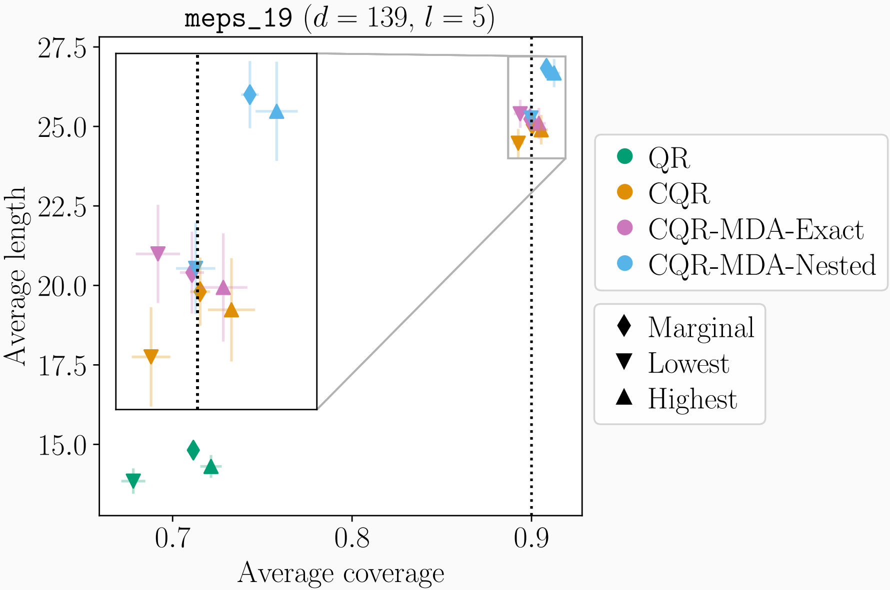
Semi-synthetic experiments

Bio data set

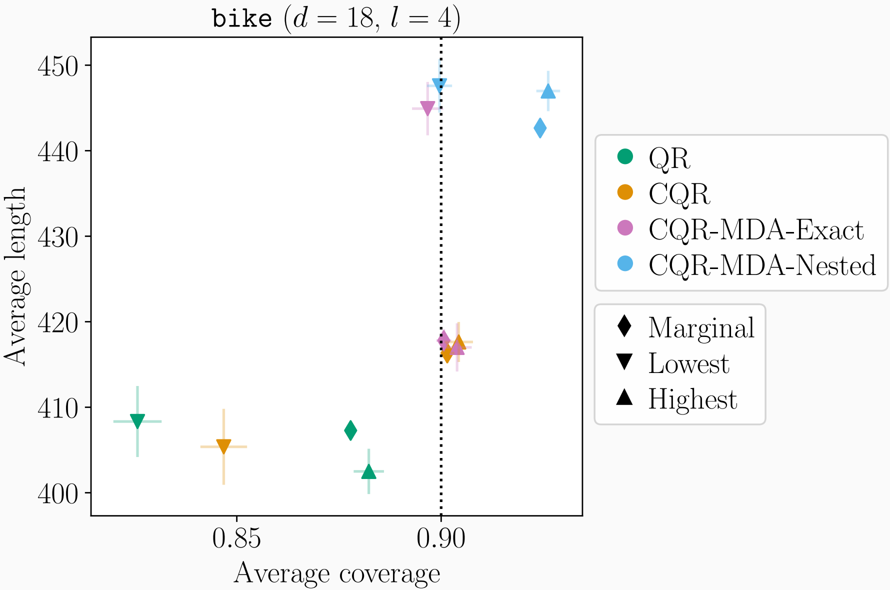


bio ($d = 9$, $l = 9$)
Average length
Average coverage
20
19
18
17
16
15
14
0.7
0.8
0.9
QR
CQR
CQR-MDA-Exact
CQR-MDA-Nested
Marginal
Lowest
Highest

`Meps_19` data set



Bike data set



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).