Conformal Prediction with Missing Values

Margaux Zaffran

54èmes Journées de Statistiques, Bruxelles, 2023
Session MALIA

Aymeric Dieuleveut
Ecole
Polytechnique
Paris - France

Julie Josse
INRIA
IDESP
Montpellier - France

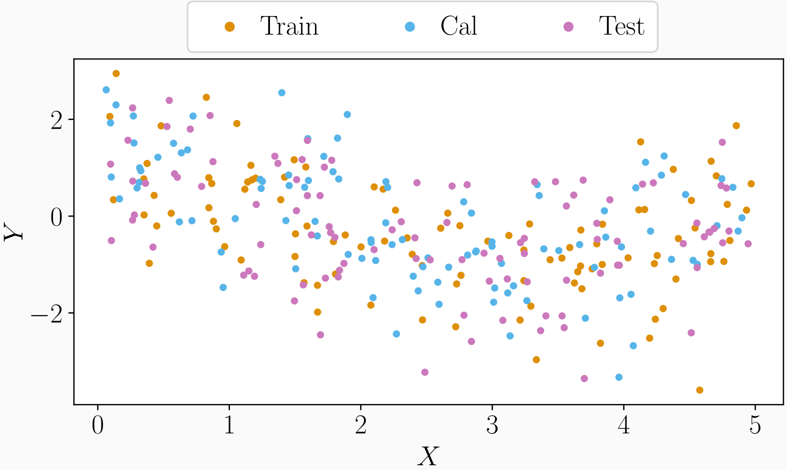
Yaniv Romano
Technion - Israel Institute
of Technology
Haifa - Israel

What about splitting the data?

- Standard Split Conformal Prediction for Mean-Regression
- Conformalized Quantile Regression

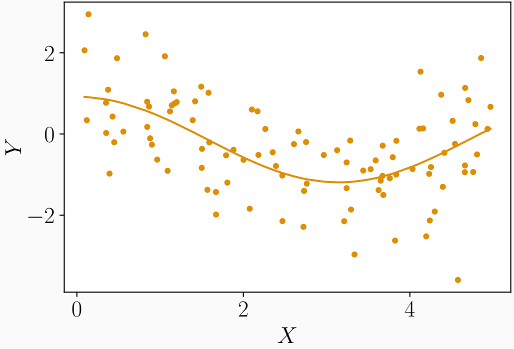
Predictive uncertainty quantification with missing values

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

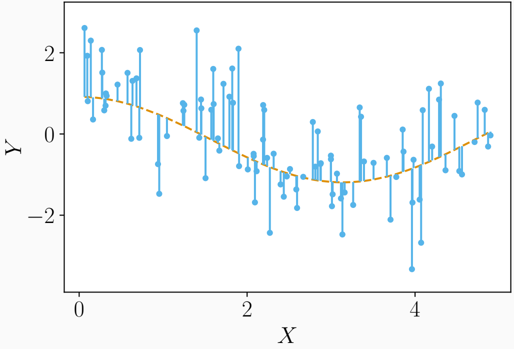
Split Conformal Prediction (SCP)[1,2,3]: training step



- 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

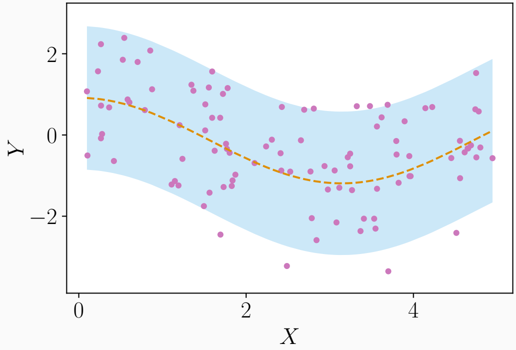
Split Conformal Prediction (SCP)[1,2,3]: calibration step



- Predict with $\hat{\mu}$
- Get the $|\text{residuals}|$, a.k.a. scores $\left\{S^{(k)}\right\}_{k\in\mathrm{Cal}}$
- 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:

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

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

Definition (Exchangeability)

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

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

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

Toy case: $Z^{(1)}$ and $Z^{(2)}$ are exchangeable if $(Z^{(1)}, Z^{(2)}) \overset{\mathcal{L}}{=} (Z^{(2)}, Z^{(1)})$.

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:

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

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

Examples of exchangeable sequences

- i.i.d. samples

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:*

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

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:*

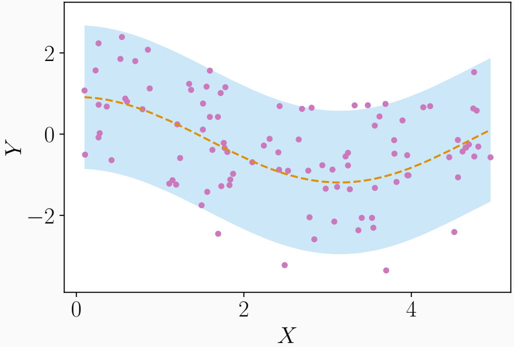
$$\mathbb{P}\left\{Y^{(n+1)} \in \widehat{C}_\alpha\left(X^{(n+1)}\right)\right\} \geq 1-\alpha.$$

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

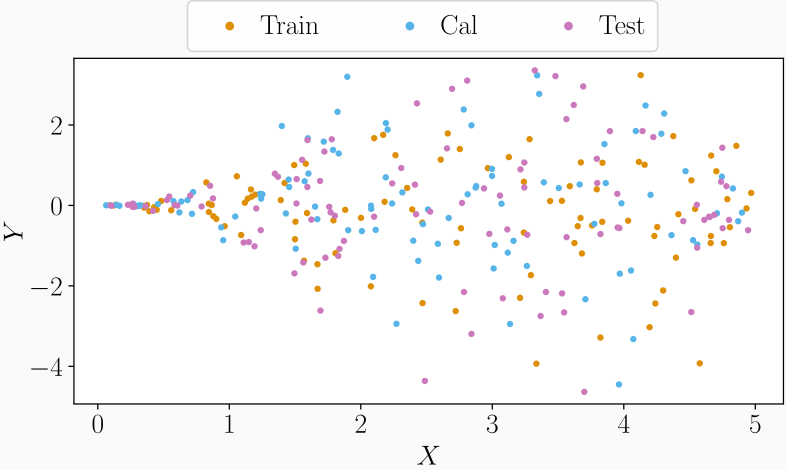
What about splitting the data?

Standard Split Conformal Prediction for Mean-Regression

Conformalized Quantile Regression

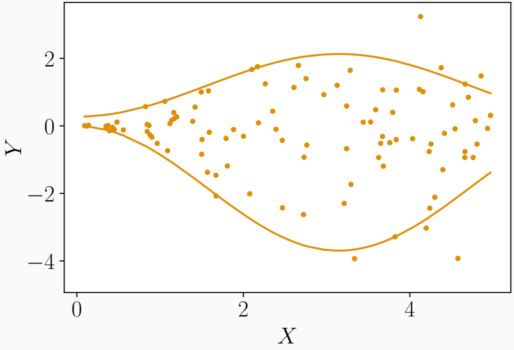
Predictive uncertainty quantification with missing values

Conformalized Quantile Regression (CQR)[4]



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

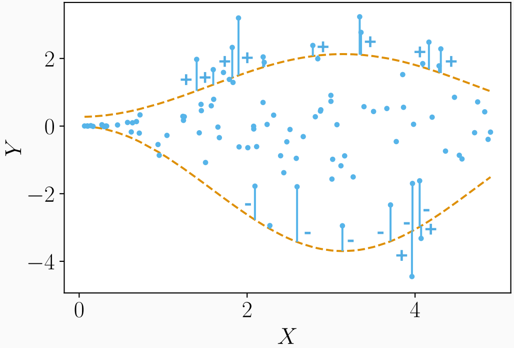
Conformalized Quantile Regression (CQR)[4]: training step



- 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

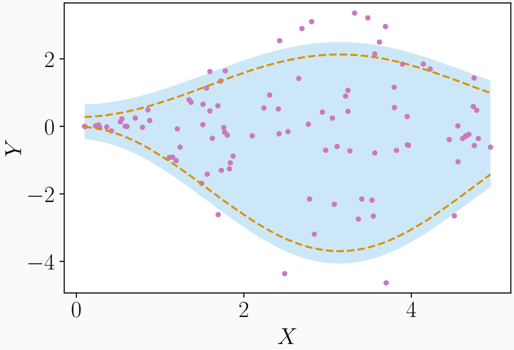


- Predict with $\widehat{QR}_{\text{lower}}$ and $\widehat{QR}_{\text{upper}}$
- Get the scores $\mathcal{S} = \left\{S^{(k)}\right\}_{\text{Cal}} \cup \{+\infty\}$
- Compute the $(1-\alpha)$ empirical quantile of $\mathcal{S}$, noted $q_{1-\alpha}(\mathcal{S})$

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

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

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 is a **particular case of SCP**.

CQR: theoretical guarantees

CQR is a **particular case of SCP**.

Therefore, it enjoys finite sample guarantees proved in Romano et al. (2019).

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 is a **particular case of SCP**.

Therefore, it enjoys finite sample guarantees proved in Romano et al. (2019).

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

Missing values are ubiquitous and challenging

Data: $\left(X^{(k)}, \qquad Y^{(k)}\right)_{k=1}^{n}$

Y	X_1	X_2	X_3
22.42	0.55	0.67	0.03
8.26	0.72	0.18	0.55
19.41	0.60	0.58	NA
19.75	0.54	0.43	0.96
7.32	NA	0.19	NA
13.55	0.65	0.69	0.50
20.75	NA	NA	0.61
9.26	0.89	NA	0.84
9.68	0.963	0.45	0.65

Missing values are ubiquitous and challenging

Data: $\left(X^{(k)}, M^{(k)}, Y^{(k)}\right)_{k=1}^{n}$

Y	X_1	X_2	X_3
22.42	0.55	0.67	0.03
8.26	0.72	0.18	0.55
19.41	0.60	0.58	NA
19.75	0.54	0.43	0.96
7.32	NA	0.19	NA
13.55	0.65	0.69	0.50
20.75	NA	NA	0.61
9.26	0.89	NA	0.84
9.68	0.963	0.45	0.65

Mask $M =$

$(M_1$	M_2	$M_3)$
0	0	0
0	0	0
0	0	1
0	0	0
1	0	1
0	0	0
1	1	0
0	1	0
0	0	0

Missing values are ubiquitous and challenging

Data: $\left(X^{(k)}, M^{(k)}, Y^{(k)}\right)_{k=1}^{n}$

Y	X_1	X_2	X_3
22.42	0.55	0.67	0.03
8.26	0.72	0.18	0.55
19.41	0.60	0.58	NA
19.75	0.54	0.43	0.96
7.32	NA	0.19	NA
13.55	0.65	0.69	0.50
20.75	NA	NA	0.61
9.26	0.89	NA	0.84
9.68	0.963	0.45	0.65

Mask $M =$

$(M_1$	M_2	$M_3)$
0	0	0
0	0	0
0	0	1
0	0	0
1	0	1
0	0	0
1	1	0
0	1	0
0	0	0

$\hookrightarrow$ 2^d potential masks.

Missing values are ubiquitous and challenging

Data: $\left(X^{(k)}, M^{(k)}, Y^{(k)}\right)_{k=1}^{n}$

Y	X_1	X_2	X_3
22.42	0.55	0.67	0.03
8.26	0.72	0.18	0.55
19.41	0.60	0.58	NA
19.75	0.54	0.43	0.96
7.32	NA	0.19	NA
13.55	0.65	0.69	0.50
20.75	NA	NA	0.61
9.26	0.89	NA	0.84
9.68	0.963	0.45	0.65

Mask $M =$

$(M_1$	M_2	$M_3)$
0	0	0
0	0	0
0	0	1
0	0	0
1	0	1
0	0	0
1	1	0
0	1	0
0	0	0

$\hookrightarrow$ 2^d potential masks.

$\hookrightarrow$ M can depend on X or Y.

Missing values are ubiquitous and challenging

Data: $\left(X^{(k)}, M^{(k)}, Y^{(k)}\right)_{k=1}^{n}$

Y	X_1	X_2	X_3
22.42	0.55	0.67	0.03
8.26	0.72	0.18	0.55
19.41	0.60	0.58	NA
19.75	0.54	0.43	0.96
7.32	NA	0.19	NA
13.55	0.65	0.69	0.50
20.75	NA	NA	0.61
9.26	0.89	NA	0.84
9.68	0.963	0.45	0.65

Mask $M =$

$(M_1$	M_2	$M_3)$
0	0	0
0	0	0
0	0	1
0	0	0
1	0	1
0	0	0
1	1	0
0	1	0
0	0	0

$\hookrightarrow$ 2^d potential masks.

$\hookrightarrow$ M can depend on X or Y.

$\Rightarrow$ Statistical and computational challenges.

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 (e.g. the mean), noted ϕ.

$$\begin{array}{c} x^{(1)} \\ x^{(2)} \\ x^{(3)} \\ x^{(4)} \end{array} \begin{bmatrix} -1 & -10 & 6 & 0 \\ 4 & \texttt{NA} & -2 & 2 \\ 5 & 1 & 2 & \texttt{NA} \\ 0 & \texttt{NA} & \texttt{NA} & 1 \end{bmatrix} \xrightarrow{\quad \phi \quad} \begin{array}{c} u^{(1)} \\ u^{(2)} \\ u^{(3)} \\ u^{(4)} \end{array} \begin{bmatrix} -1 & -10 & 6 & 0 \\ 4 & -4.5 & -2 & 2 \\ 5 & 1 & 2 & 1 \\ 0 & -4.5 & 3 & 1 \end{bmatrix}$$

Supervised learning with missing values: impute-then-regress

Impute-then-regress procedures are widely used.

1. Replace NA using an imputation function (e.g. the mean), noted ϕ.

$$
\begin{array}{c|c|c|c|c|}
x^{(1)} & -1 & -10 & 6 & 0 \\
x^{(2)} & 4 & \mathtt{NA} & -2 & 2 \\
x^{(3)} & 5 & 1 & 2 & \mathtt{NA} \\
x^{(4)} & 0 & \mathtt{NA} & \mathtt{NA} & 1
\end{array}
\xrightarrow{\phi}
\begin{array}{c|c|c|c|c|}
u^{(1)} & -1 & -10 & 6 & 0 \\
u^{(2)} & 4 & -4.5 & -2 & 2 \\
u^{(3)} & 5 & 1 & 2 & 1 \\
u^{(4)} & 0 & -4.5 & 3 & 1
\end{array}
$$

2. Train your algorithm (Random Forest, Neural Nets, etc.) on the imputed data: $\left\{ \underbrace{\phi\left(X^{(k)}_{\mathrm{obs}(M^{(k)})}, M^{(k)}\right)}_{U^{(k)} = \text{imputed } X^{(k)}}, Y^{(k)} \right\}_{k=1}^{n}$.

Supervised learning with missing values: impute-then-regress

Impute-then-regress procedures are widely used.

1. Replace NA using an imputation function (e.g. the mean), noted ϕ.

$$
\begin{array}{c}
x^{(1)} \\ x^{(2)} \\ x^{(3)} \\ x^{(4)}
\end{array}
\begin{array}{|c|c|c|c|}
\hline
-1 & -10 & 6 & 0 \\ \hline
4 & \texttt{NA} & -2 & 2 \\ \hline
5 & 1 & 2 & \texttt{NA} \\ \hline
0 & \texttt{NA} & \texttt{NA} & 1 \\ \hline
\end{array}
\xrightarrow{\quad\phi\quad}
\begin{array}{c}
u^{(1)} \\ u^{(2)} \\ u^{(3)} \\ u^{(4)}
\end{array}
\begin{array}{|c|c|c|c|}
\hline
-1 & -10 & 6 & 0 \\ \hline
4 & -4.5 & -2 & 2 \\ \hline
5 & 1 & 2 & 1 \\ \hline
0 & -4.5 & 3 & 1 \\ \hline
\end{array}
$$

2. Train your algorithm (Random Forest, Neural Nets, etc.) on the imputed data: $\left\{ \underbrace{\phi\left(X^{(k)}_{\mathrm{obs}(M^{(k)})}, M^{(k)}\right)}_{U^{(k)}=\text{imputed } X^{(k)}}, Y^{(k)} \right\}_{k=1}^{n}$.

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

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)

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

For example: $\alpha = 0.1$ and obtain a 90% coverage interval.

Predictive uncertainty quantification with missing values

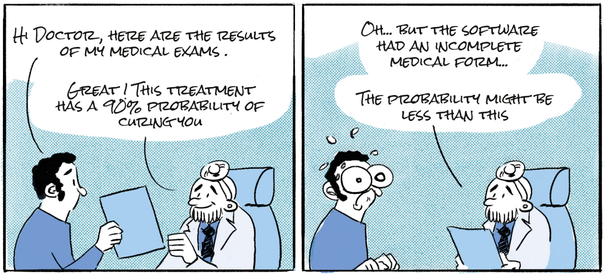
Goal: predict $Y^{(n+1)}$ with **confidence** $1-\alpha$, i.e. build the smallest $\mathcal{C}_\alpha$ such that:

1. Marginal Validity (MV)

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

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 @theoremlinger

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)

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

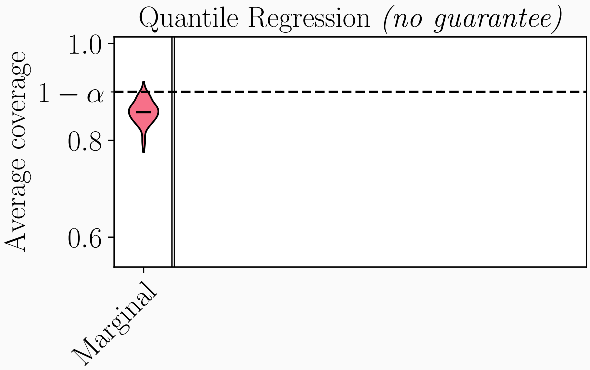
2. Mask-Conditional-Validity (MCV)

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

3 considered approaches to reach these goals.

	Quantile Regression (QR)		
(MV)	?		
(MCV)	?		

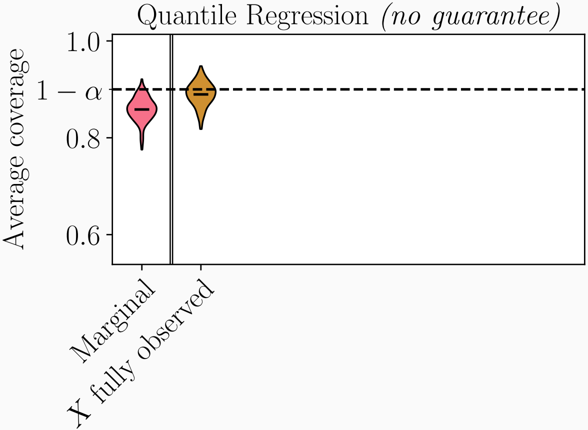
Quantile Regression (QR) intervals



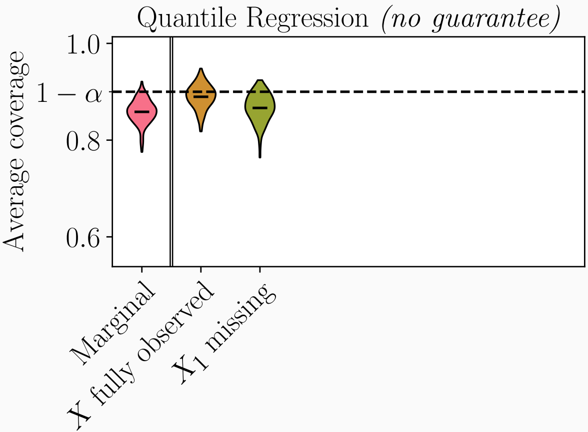
- Marginal validity (eq. (MV), i.e. on average) is not reached!

	QR		
(MV)	✗		
(MCV)			

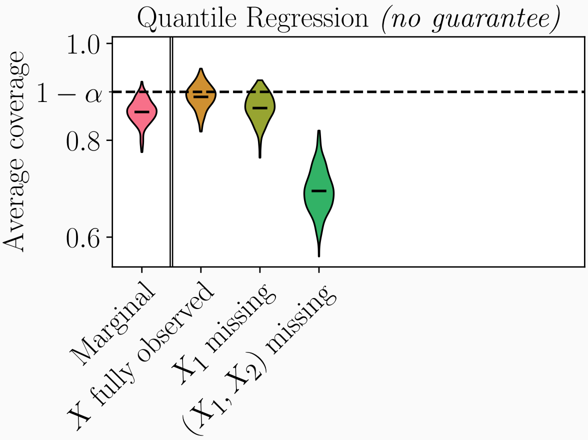
Quantile Regression (QR) intervals



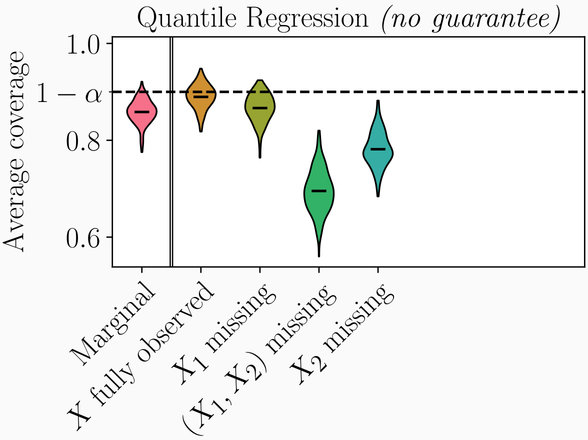
Quantile Regression (QR) intervals



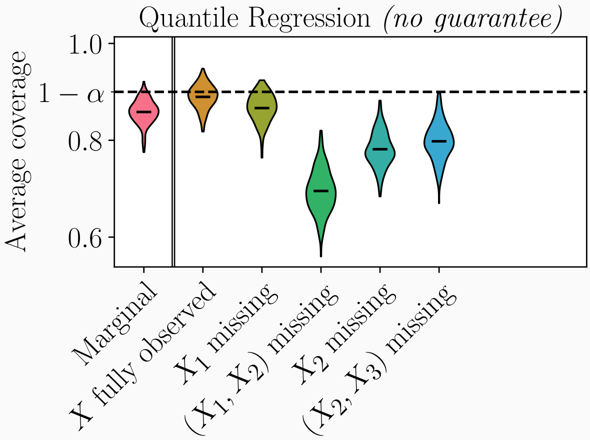
Quantile Regression (QR) intervals



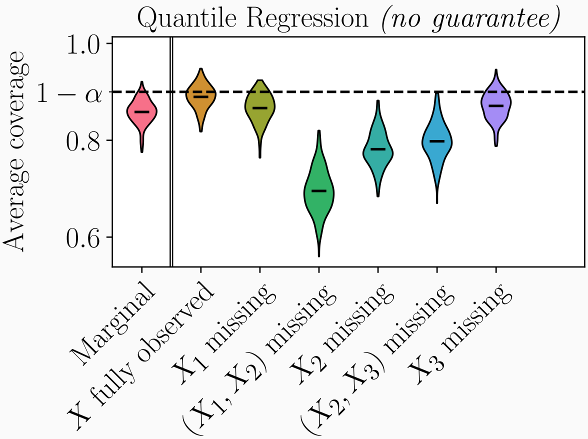
Quantile Regression (QR) intervals



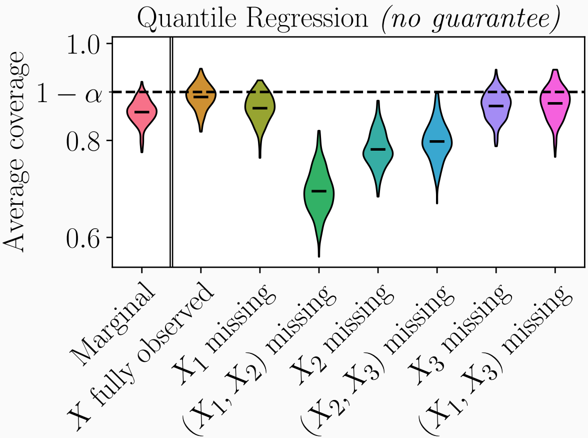
Quantile Regression (QR) intervals



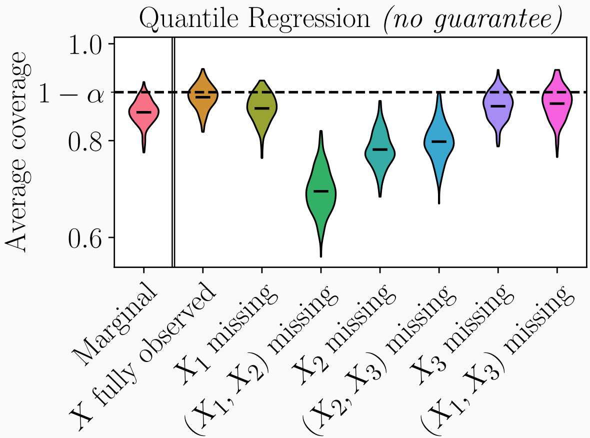
Quantile Regression (QR) intervals



Quantile Regression (QR) intervals



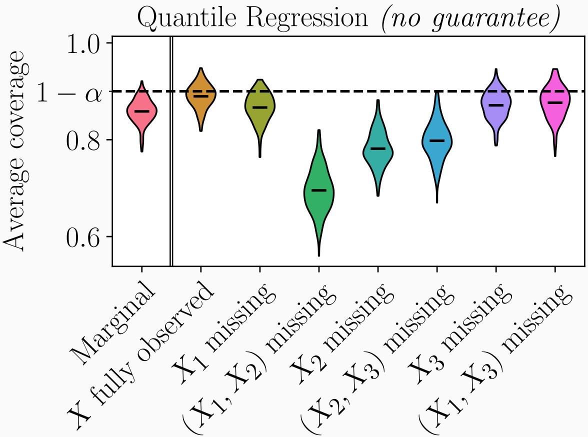
Quantile Regression (QR) intervals



- The predictive uncertainty strongly depends on the mask

	QR		
(MV)	✗		
(MCV)	✗		

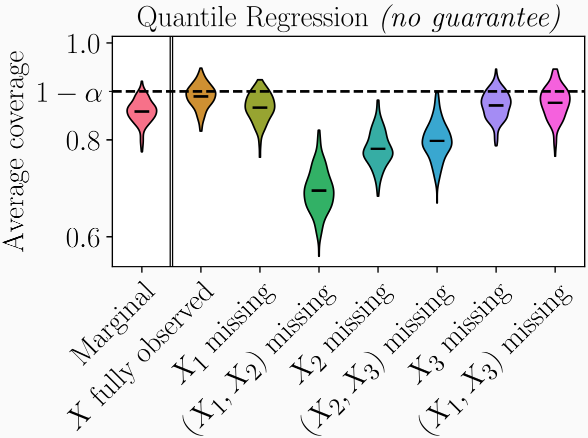
Quantile Regression (QR) intervals



- The predictive uncertainty strongly depends on the mask

$\hookrightarrow$ missing values induce heteroskedasticity

Quantile Regression (QR) intervals



- The predictive uncertainty strongly depends on the mask

$\hookrightarrow$ missing values induce heteroskedasticity

$\hookrightarrow$ supported by theory on the Gaussian Linear Model

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

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.)

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

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.)

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

- Even with an homoskedastic noise, missingness generates heteroskedasticity

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.)

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

- 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)**

CP is marginally valid (MV) after imputation

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[1] ϕ*:*

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

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

CP is marginally valid (MV) after imputation

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[1] ϕ*:*

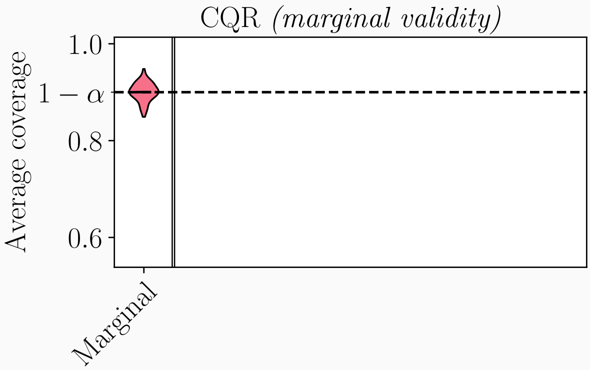
$\left(\phi\left(X^{(k)}_{obs(M^{(k)})}, M^{(k)}\right), Y^{(k)}\right)_{k=1}^{n}$ *are* ***exchangeable****.*

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

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

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

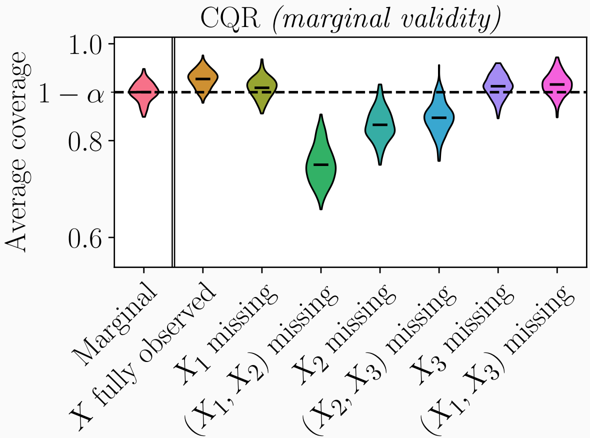
CQR is marginally valid on imputed data sets



- Marginal (i.e. average) coverage is indeed recovered!

	QR	CQR
(MV)	✗	✓
(MCV)	✗	

CQR is marginally valid on imputed data sets

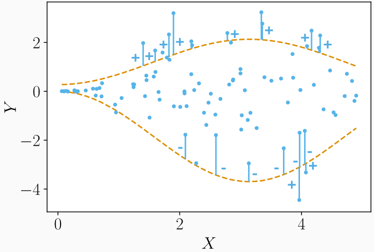


- Disparities between masks is not corrected by the conformalization step.

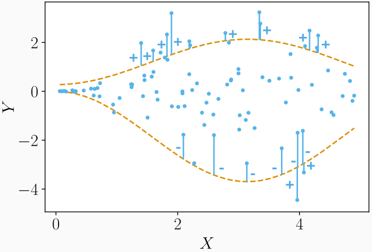
	QR	CQR
(MV)	✗	✓
(MCV)	✗	✗

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!

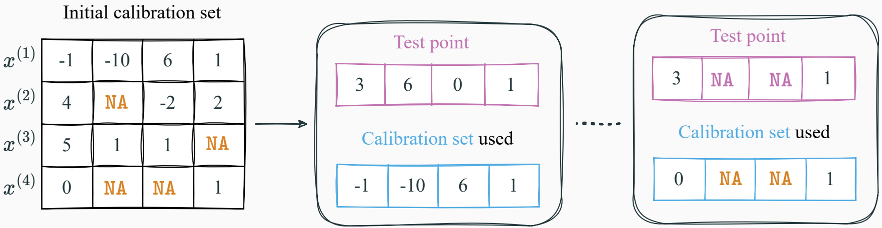


Observation: the α-correction term is computed among all the data points, regardless of their mask!



Warning: 2^d possible masks

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



Missing Data Augmentation (MDA)

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

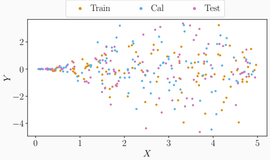
$x^{(1)}$	-1	-10	6	1
$x^{(2)}$	4	NA	-2	2
$x^{(3)}$	5	1	1	NA
$x^{(4)}$	0	NA	NA	1

Calibration set used

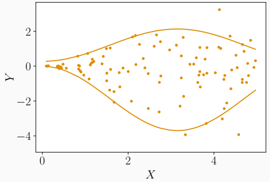
$\tilde{x}^{(1)}$	-1	NA	NA	1
$\tilde{x}^{(2)}$	4	NA	NA	2
$\tilde{x}^{(3)}$				
$\tilde{x}^{(4)}$	0	NA	NA	1

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

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

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 $(X^{(n+1)}, M^{(n+1)})$:

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

   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`

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

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 $(X^{(n+1)}, M^{(n+1)})$:

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

   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 $\left(X^{(n+1)}, M^{(n+1)}\right)$:

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

   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)

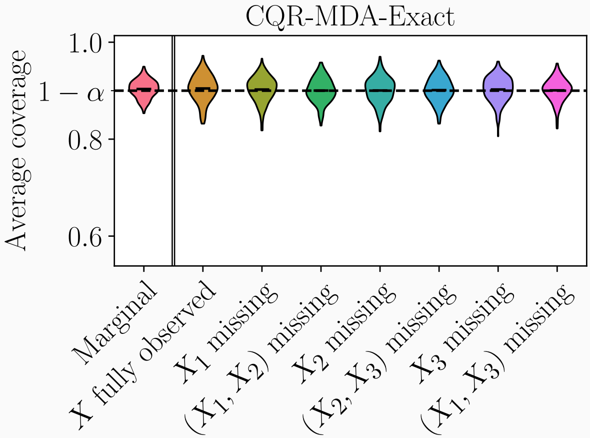
If the data is exchangeable and $M \perp\!\!\!\perp (X, Y)$, then for almost all imputation function CP-MDA-Exact is such that for any $m \in \{0,1\}^d$:

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

and if additionally the scores are almost surely distinct:

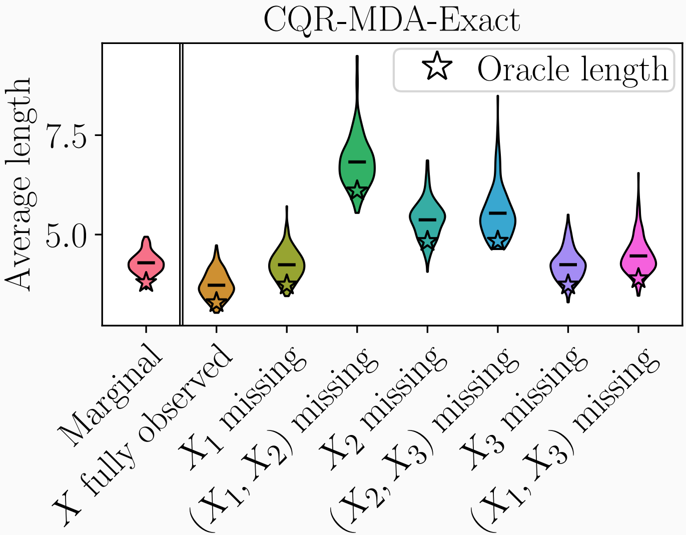
$$\mathbb{P}\left(Y \in \widehat{C}_\alpha(X, m) \,|\, M = m\right) \leq 1 - \alpha + \frac{1}{1 + \#\mathrm{Cal}^{\mathrm{m}}}.$$

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



	QR	CQR	CQR-MDA
(MV)	✗	✓	✓
(MCV)	✗	✗	✓

MDA achieves Mask-Conditional-Validity in an informative way



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

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:

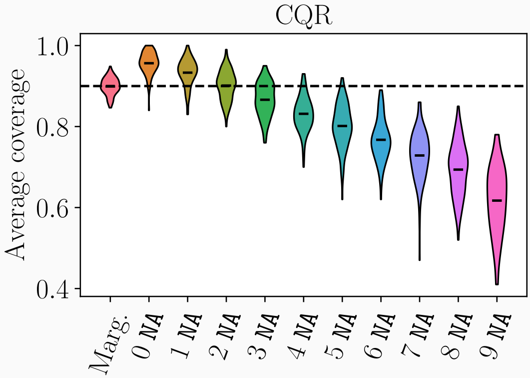
Some settings

- 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%

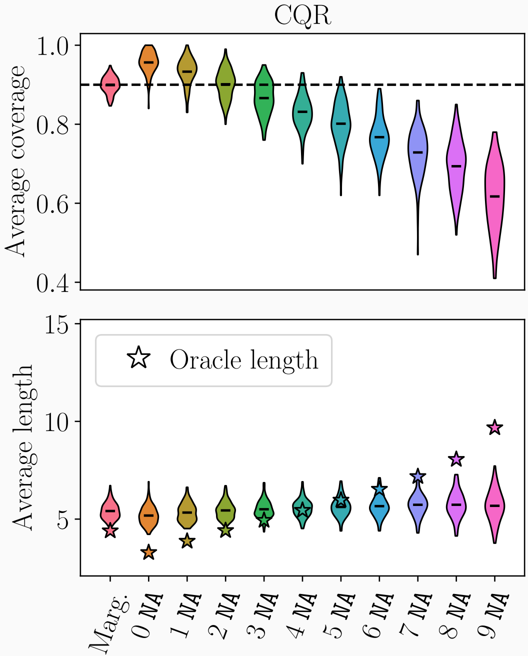
Some settings

- 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

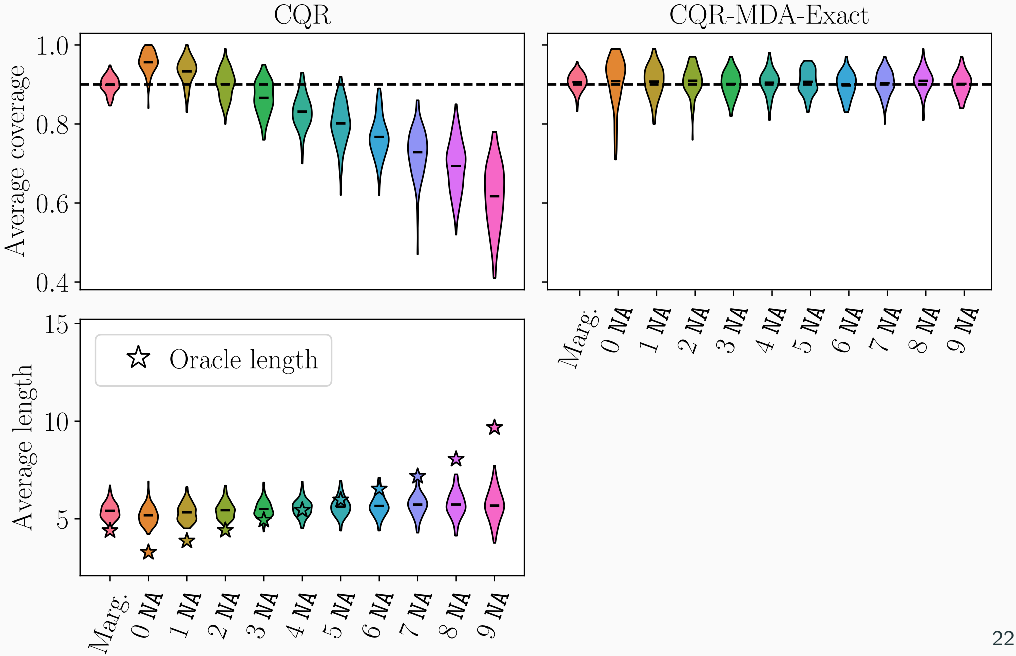
Synthetic experiments (Gaussian linear model, $d = 10$)



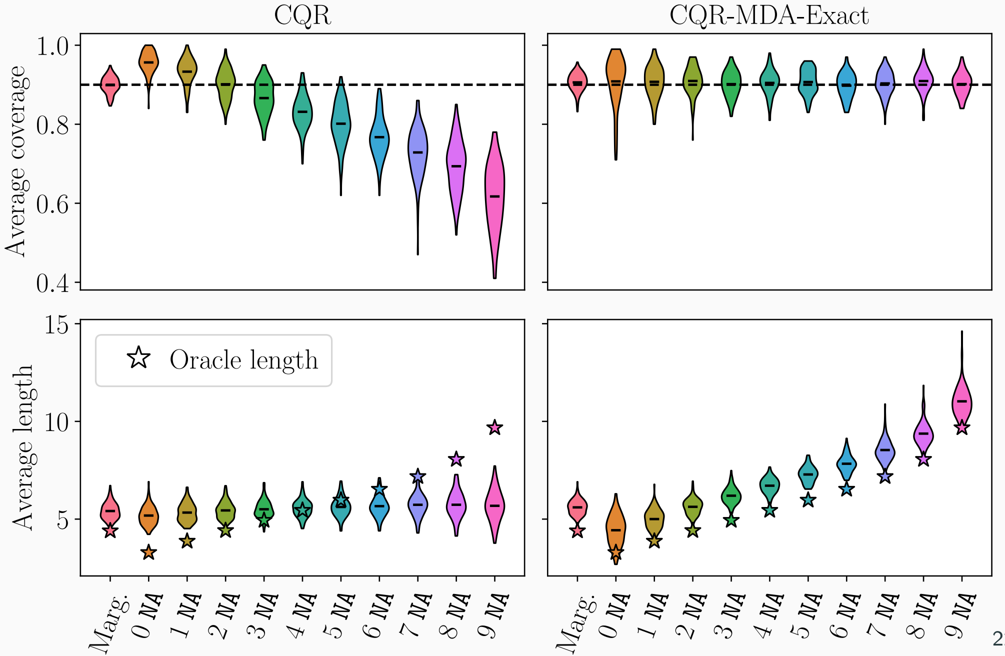
Synthetic experiments (Gaussian linear model, $d = 10$)



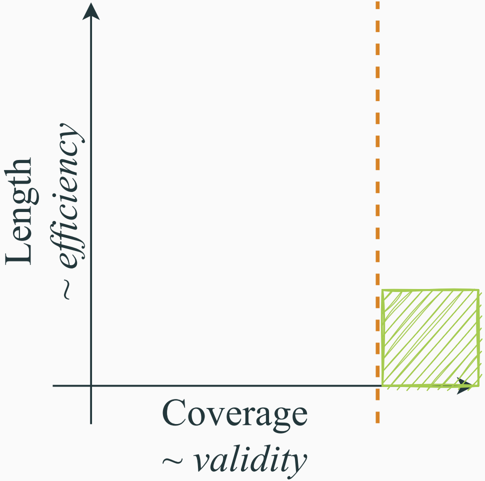
Synthetic experiments (Gaussian linear model, $d = 10$)



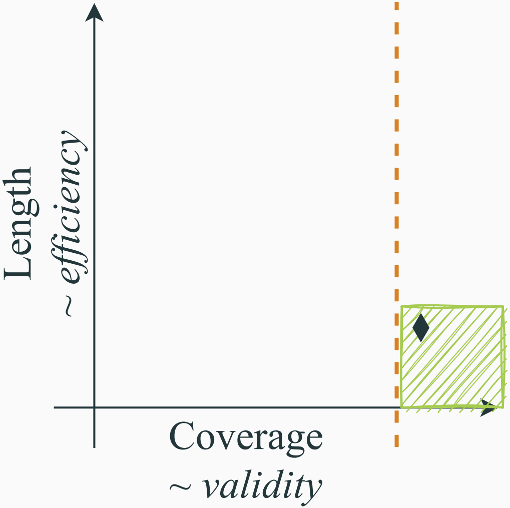
Synthetic experiments (Gaussian linear model, $d = 10$)



Before more experiments, visualisation



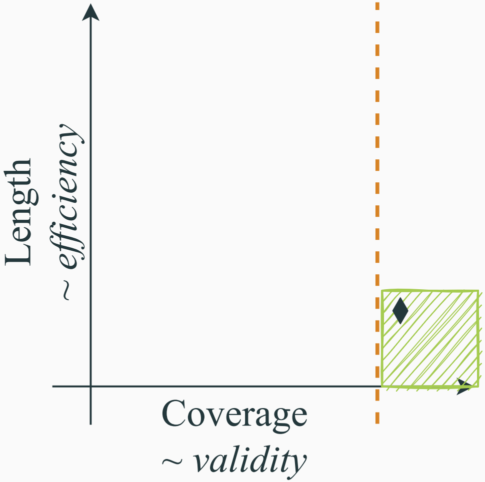
Before more experiments, visualisation



♦ : marginal coverage, i.e.

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

Before more experiments, visualisation



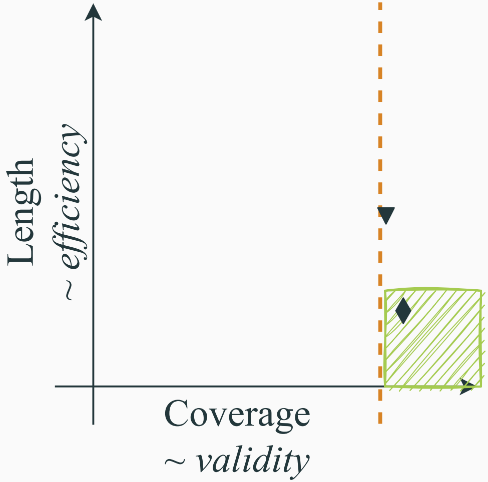
- ♦ : marginal coverage, i.e.

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

- ▼ : lowest coverage, i.e.

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

Before more experiments, visualisation



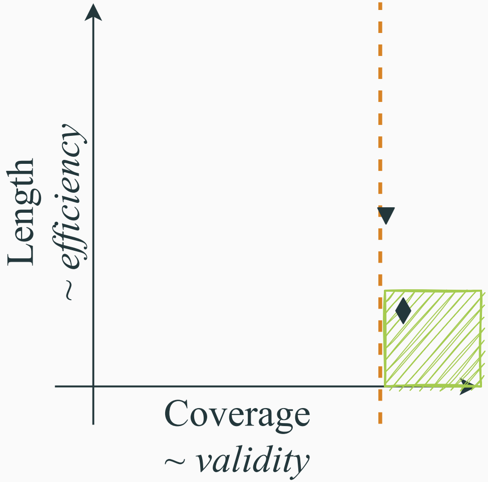
- ♦ : marginal coverage, i.e.

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

- ▼ : lowest coverage, i.e.

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

Before more experiments, visualisation



- ♦ : marginal coverage, i.e.

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

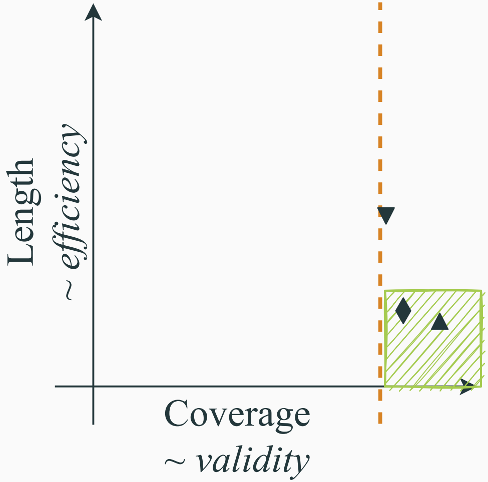
- ▼ : lowest coverage, i.e.

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

- ▲ : highest coverage, i.e.

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

Before more experiments, visualisation



- ♦ : marginal coverage, i.e.

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

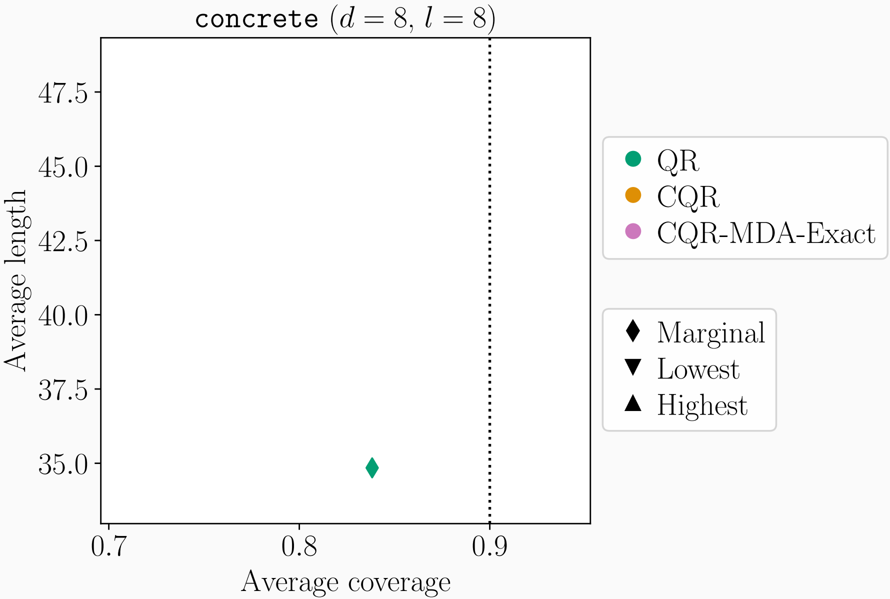
- ▼ : lowest coverage, i.e.

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

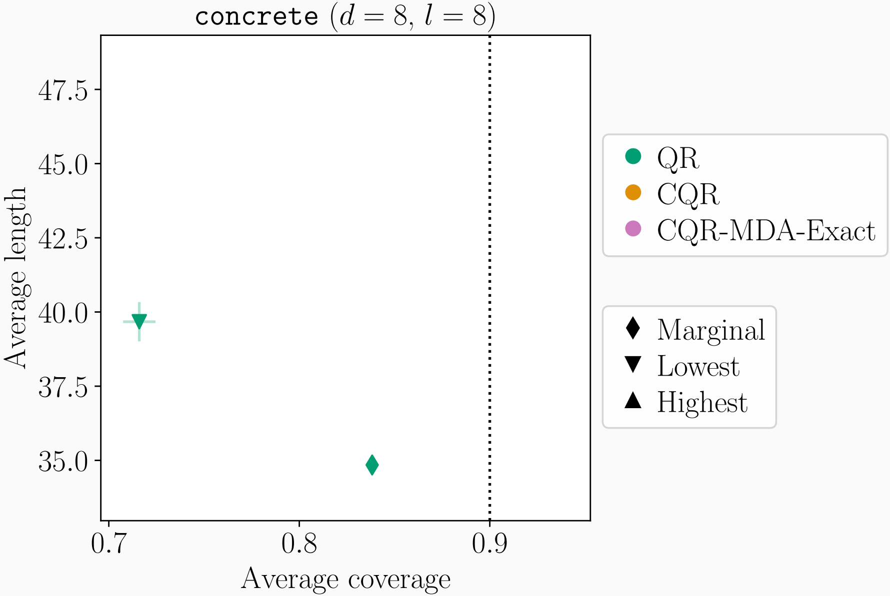
- ▲ : highest coverage, i.e.

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

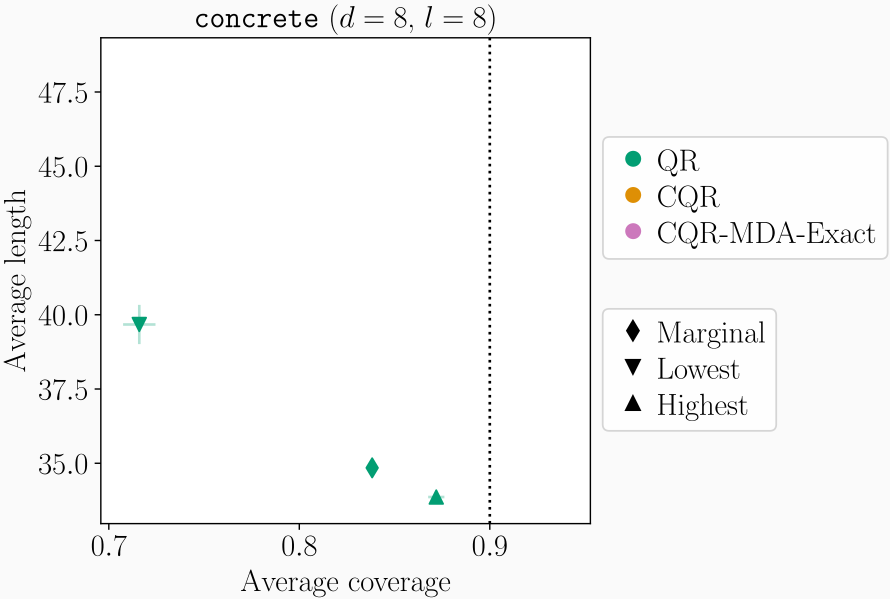
Semi-synthetic experiments



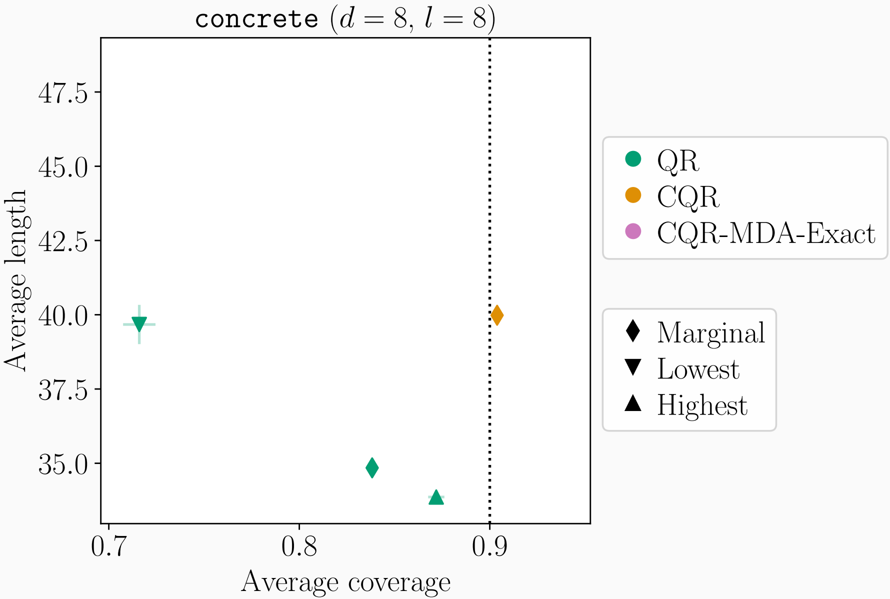
Semi-synthetic experiments



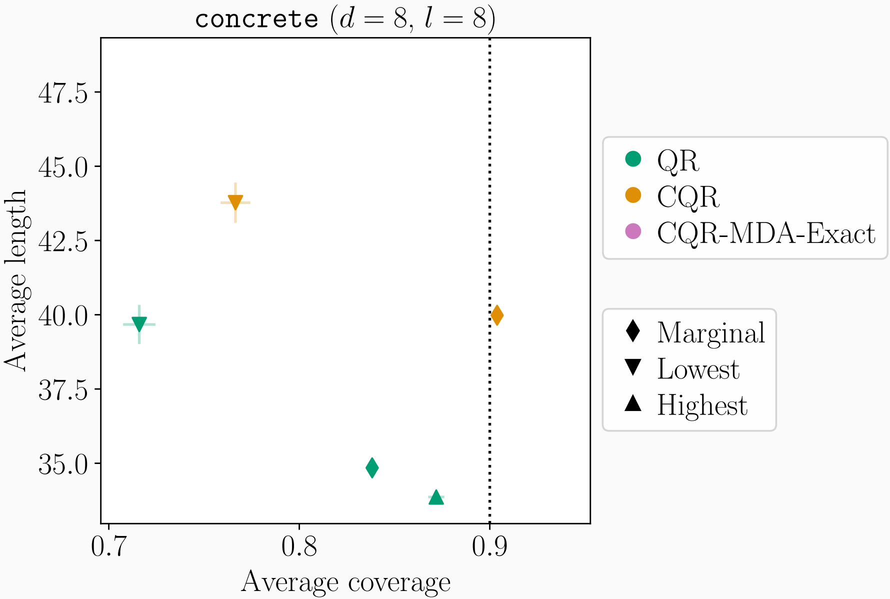
Semi-synthetic experiments



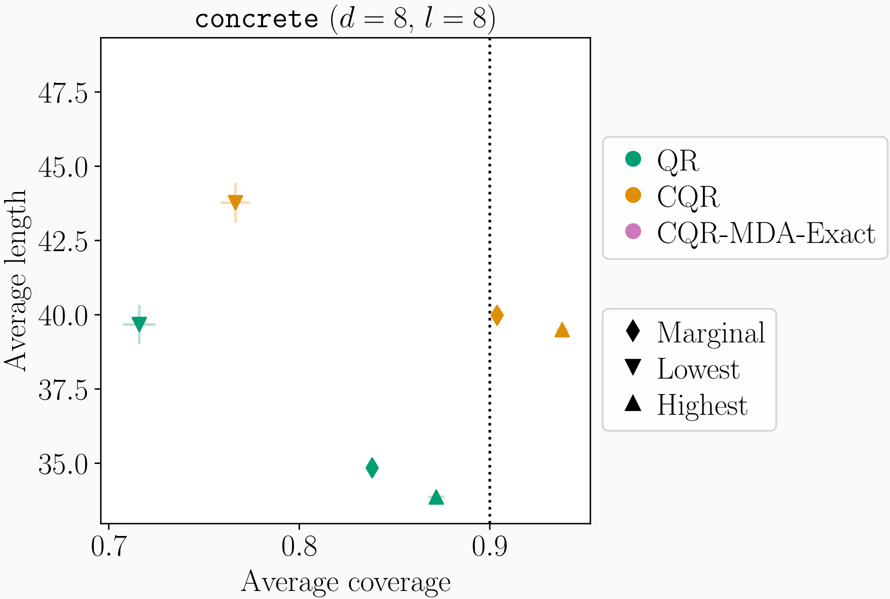
Semi-synthetic experiments



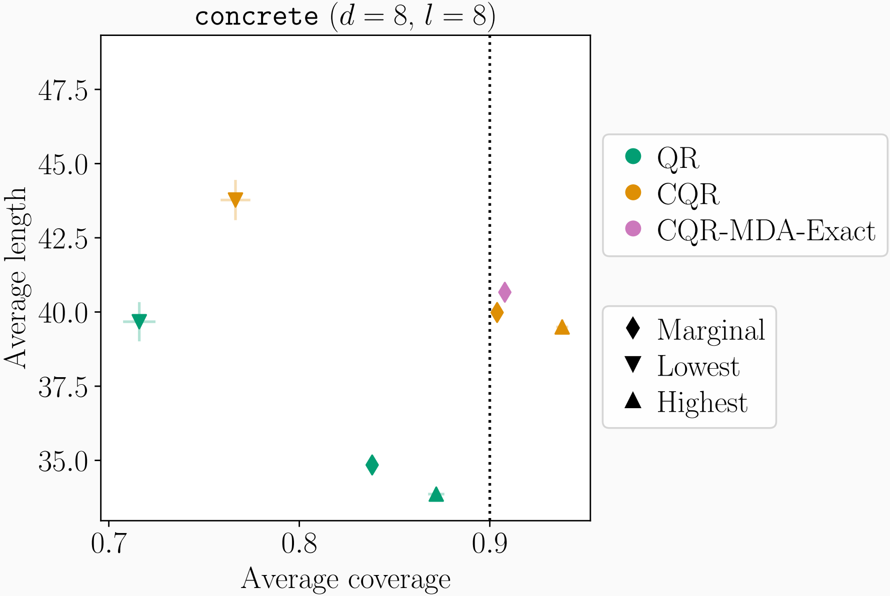
Semi-synthetic experiments



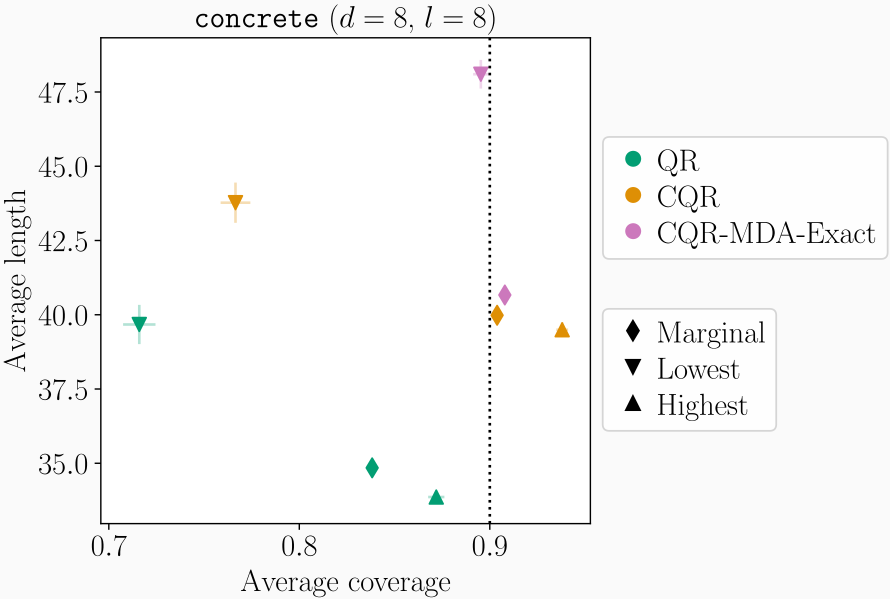
Semi-synthetic experiments



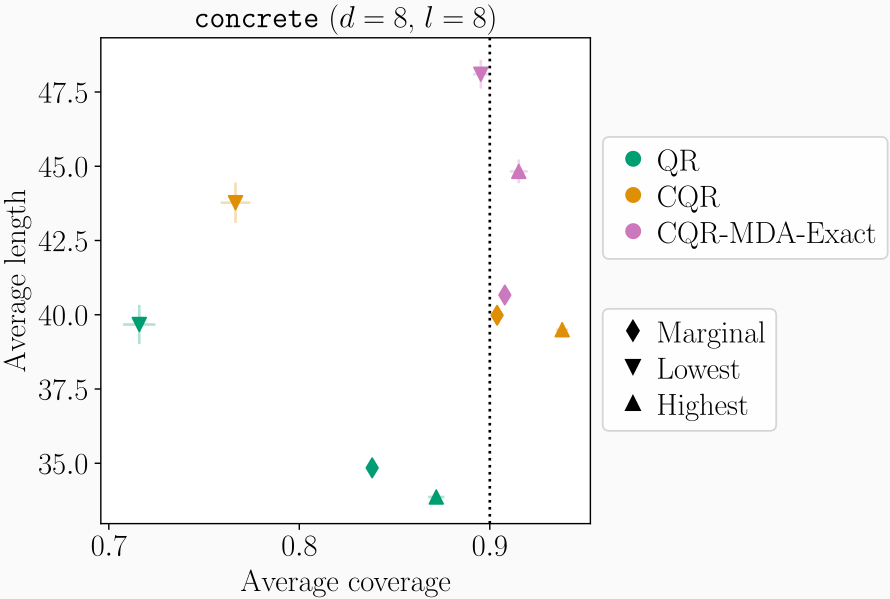
Semi-synthetic experiments



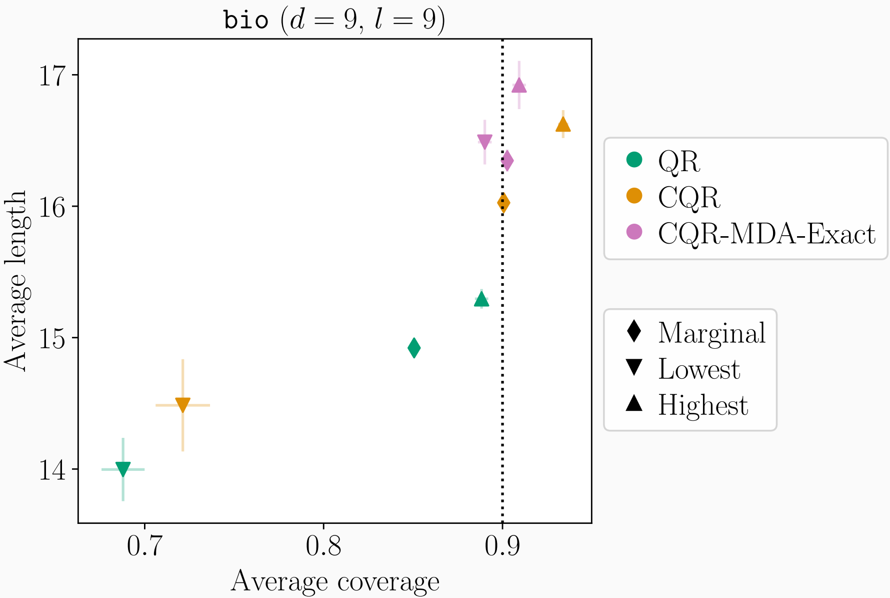
Semi-synthetic experiments



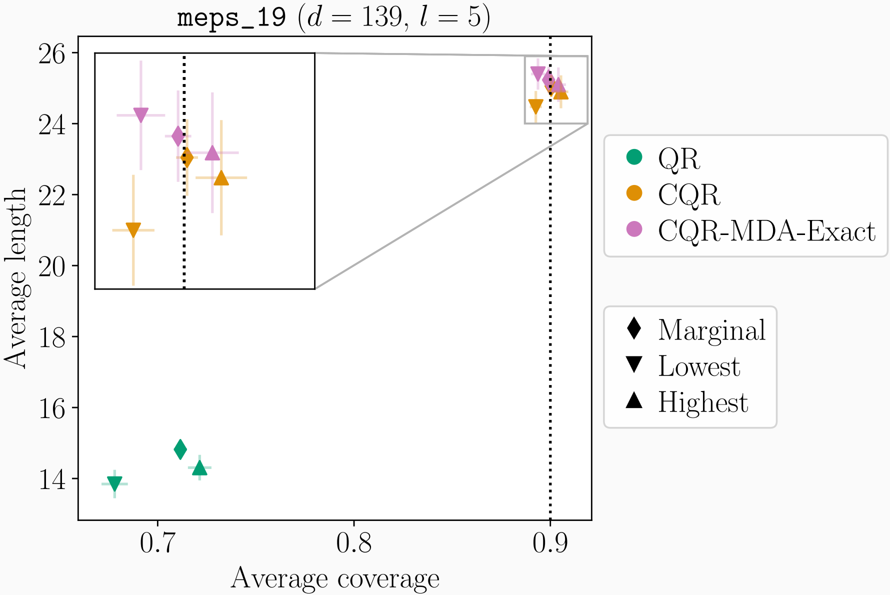
Semi-synthetic experiments



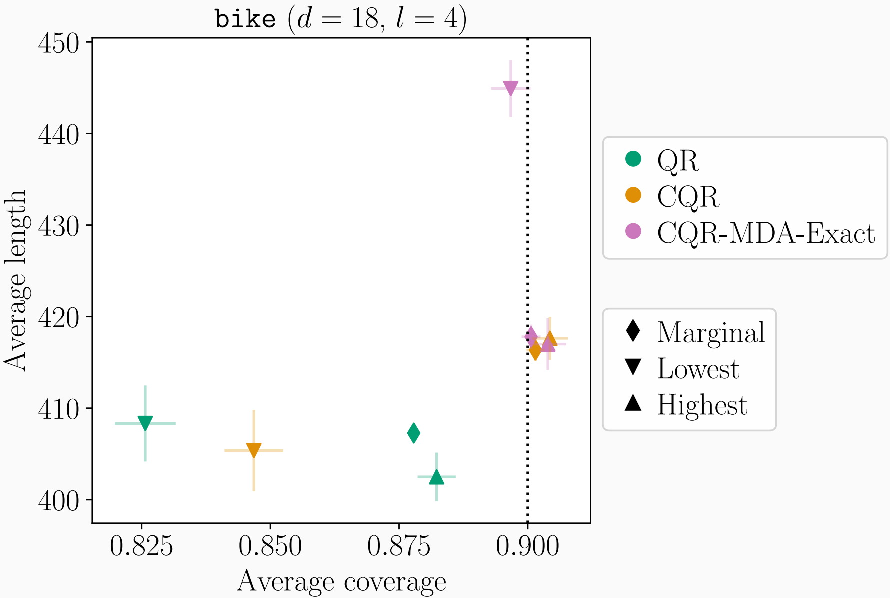
Semi-synthetic experiments



Semi-synthetic experiments



Semi-synthetic experiments



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
  $\hookrightarrow$ Many useful statistical tasks

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

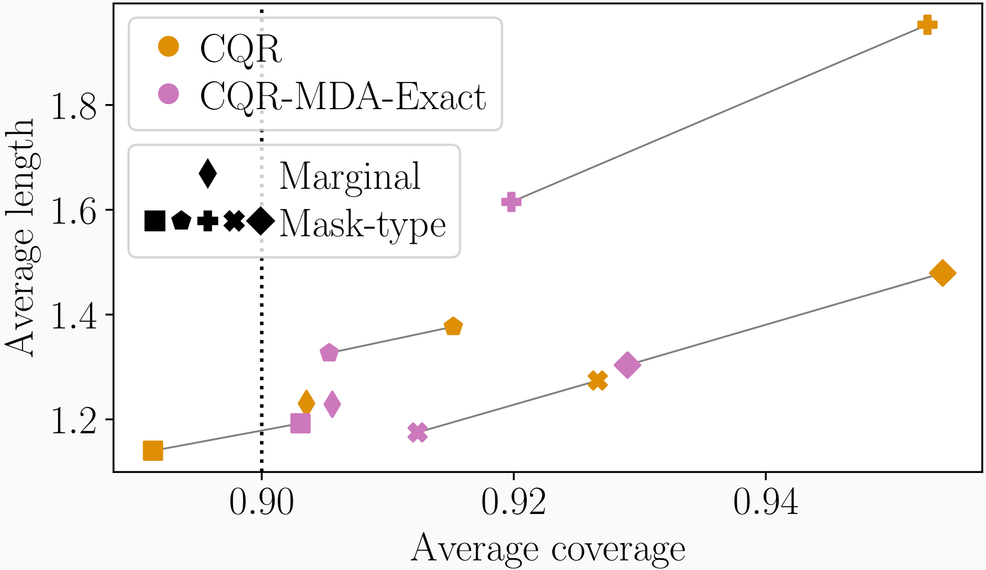
TraumaBase®: decision support for trauma patients

- 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: from 0% to 24% of missing values by features, with a total average of 7%.

Real data experiment: TraumaBase®, critical care medicine



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

Paper $\longrightarrow$

Poster $\longrightarrow$

Code $\longrightarrow$

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).

Thanks for listening! Any question? :)

References

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*.

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

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

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