Introduction to Conformal Prediction
Extension to missing values

Margaux Zaffran

MIND & SODA Seminar

June 13, 2023

Aymeric Dieuleveut
Ecole Polytechnique
Paris (France)

Julie Josse
PreMeDICaL
INRIA
Montpellier (France)

Yaniv Romano
Technion - Israel Institute of Technology
Haifa (Israel)

Setting

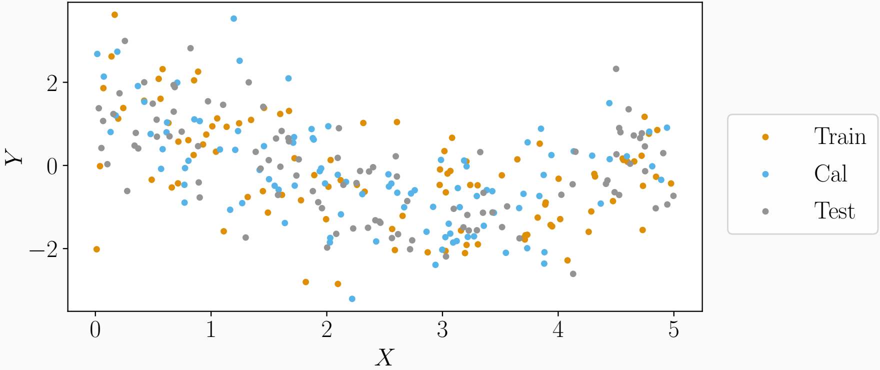
- $(X, Y) \in \mathbb{R}^d \times \mathbb{R}$ random variables
- n training samples $\left(X^{(i)}, Y^{(i)}\right)_{i=1}^n$
- Goal: predict an unseen point $Y^{(n+1)}$ at $X^{(n+1)}$ with **confidence**
- How? Given a miscoverage level $\alpha \in [0, 1]$, build a predictive set $\mathcal{C}_\alpha$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

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

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

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

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

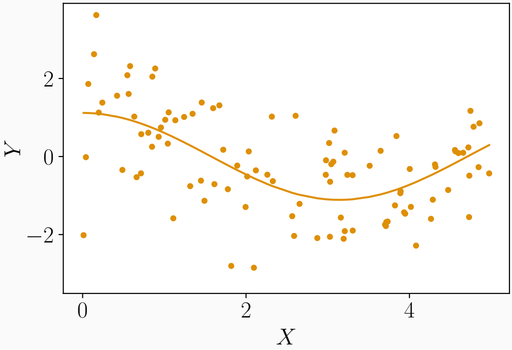


[1] Vovk et al. (2005), *Algorithmic Learning in a Random World*

[2] Papadopoulos et al. (2002), *Inductive Confidence Machines for Regression*, ECML

[3] Lei et al. (2018), *Distribution-Free Predictive Inference for Regression*, JRSS B

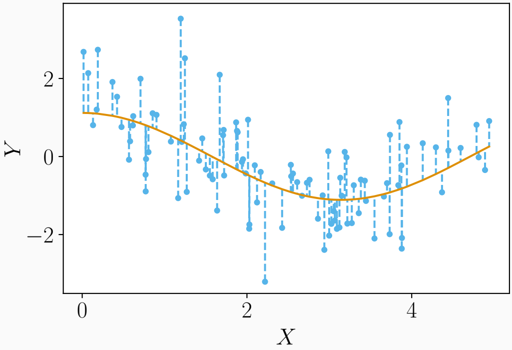
Split Conformal Prediction (SCP)[1,2,3]: training step



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

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

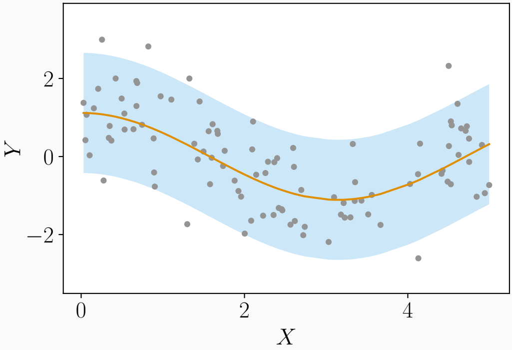
Split Conformal Prediction (SCP)[1,2,3]: calibration step



- Predict with $\hat{\mu}$
- Get the $|$residuals$|$
- Compute the $(1-\alpha)$ empirical quantile of the $|$residuals$| \cup \{+\infty\}$, noted $q_{1-\alpha}$(residuals)

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

[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

Standard mean-regression SCP: formally

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

$$\mathcal{S} = \{S^{(i)} = |\hat{A}\left(X^{(i)}\right) - Y^{(i)}|, i \in \mathrm{Cal}\} \cup \{+\infty\}$$

(+ worst-case scenario)

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

$$\widehat{C}_\alpha\left(X^{(n+1)}\right) = \left[\hat{A}\left(X^{(n+1)}\right) - q_{1-\alpha}(\mathcal{S}); \hat{A}\left(X^{(n+1)}\right) + q_{1-\alpha}(\mathcal{S})\right]$$

Definition (Exchangeability)

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

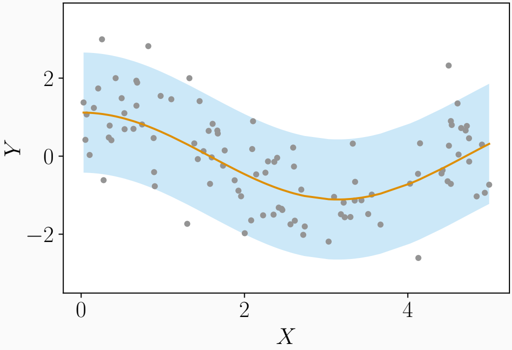
$$\mathbb{P}\left\{Y^{(n+1)} \in \widehat{C}_\alpha\left(X^{(n+1)}\right)\right\} \geq 1 - \alpha.$$

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

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

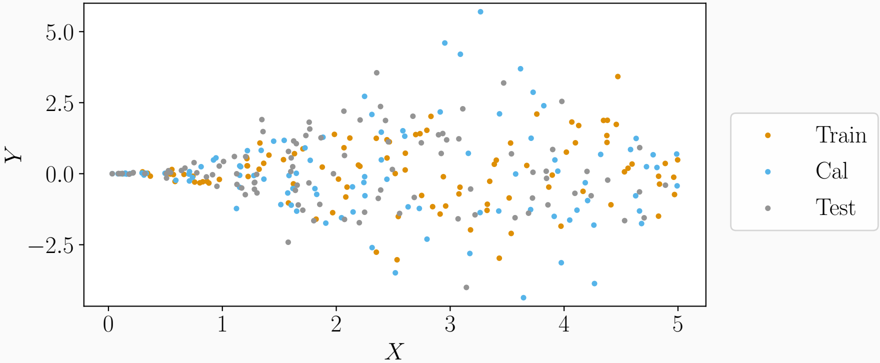
✗ Marginal coverage: $\mathbb{P}\left\{Y^{(n+1)} \in \widehat{C}_\alpha\left(X^{(n+1)}\right) \cancel{| X^{(n+1)} = x}\right\} \geq 1 - \alpha$

Standard mean-regression SCP is not adaptive



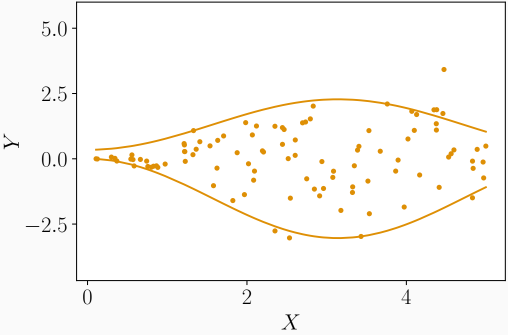
- Predict with $\hat{\mu}$
- Build $\widehat{C}_\alpha(x)$: $[\hat{\mu}(x) \pm q_{1-\alpha}(\mathcal{S})]$

Conformalized Quantile Regression (CQR)[4]



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

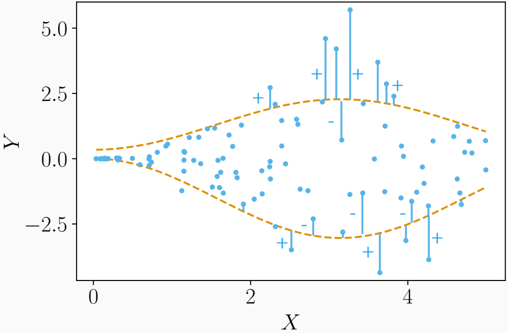
Conformalized Quantile Regression (CQR)[4]



- Learn (or get) $\widehat{QR}_{\alpha/2}$ and $\widehat{QR}_{1-\alpha/2}$

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

Conformalized Quantile Regression (CQR)[4]

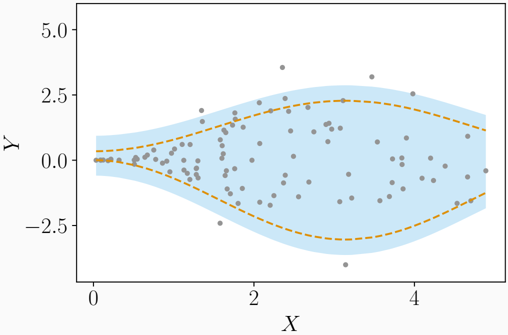


- Predict with $\widehat{QR}_{\alpha/2}$ and $\widehat{QR}_{1-\alpha/2}$
- Get the scores $\mathcal{S} = \left\{S^{(i)}\right\}_{\text{Cal}} \cup \{+\infty\}$
- Compute the $(1-\alpha)$ empirical quantile of $\mathcal{S}$, noted $q_{1-\alpha}(\mathcal{S})$

$$\hookrightarrow \quad S^{(i)} := \max\left\{\widehat{QR}_{\alpha/2}\left(X^{(i)}\right) - Y^{(i)}, Y^{(i)} - \widehat{QR}_{1-\alpha/2}\left(X^{(i)}\right)\right\}$$

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

Conformalized Quantile Regression (CQR)[4]



- Predict with $\widehat{QR}_{\alpha/2}$ and $\widehat{QR}_{1-\alpha/2}$

- Build

$$\widehat{C}_\alpha(x) = [\widehat{QR}_{\alpha/2}(x) - q_{1-\alpha}(\mathcal{S}); \widehat{QR}_{1-\alpha/2}(x) + q_{1-\alpha}(\mathcal{S})]$$

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

Generalization: SCP is defined by the conformity scores

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

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

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

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

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

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

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

SCP: theoretical guarantees generalized

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

Theorem

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

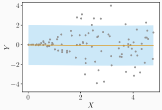
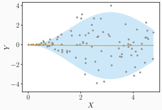
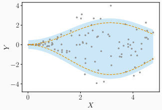
$$\mathbb{P}\left\{Y^{(n+1)} \in \widehat{C}_\alpha\left(X^{(n+1)}\right)\right\} \geq 1-\alpha.$$

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

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

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

SCP: what choices for the regression scores?

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

SCP: summary

Split conformal prediction is simple to compute and works:

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

Two interests:

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

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

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

Challenges and open directions (non-exhaustive references)

1. Providing a form of conditional guarantee
2. Tradeoffs between computational cost and statistical efficiency (i.e. variability of the estimators, *efficiency* of the predictive sets)
3. Going beyond the exchangeability assumption

CP is a very active field of research. Many developments focus on **adapting CP to specific frameworks**, such as: Survival Analysis (Candès et al., 2023), Causal Inference (Lei and Candès, 2021; Jin et al., 2023), NLP (Schuster et al., 2022), RL (Taufiq et al., 2022), applications (medical (Angelopoulos et al., 2022; Lu et al., 2022), energy (Kath and Ziel, 2021), etc.) and more.

Missing values: ubiquitous in data science practice

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

If each entry has a probability 0.01 of being missing:

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

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

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

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

Handling missing values depends on pattern and mechanism

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

Example

We observe $(\texttt{NA}, 6, 2)(-1, \texttt{NA}, 2)(-1, \texttt{NA}, \texttt{NA})$. Then
$m = (1, 0, 0) m = (0, 1, 0) m = (0, 1, 1)$.

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

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

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

Supervised learning with missing values

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

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

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

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

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

Impute-then-regress+conformalization is marginally valid

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

Lemma (Exchangeability after imp., Zaffran et al., 2023)

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

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

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

$\Rightarrow$ Conformal prediction applied on an imputed data set still enjoys marginal guarantees[9]:

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

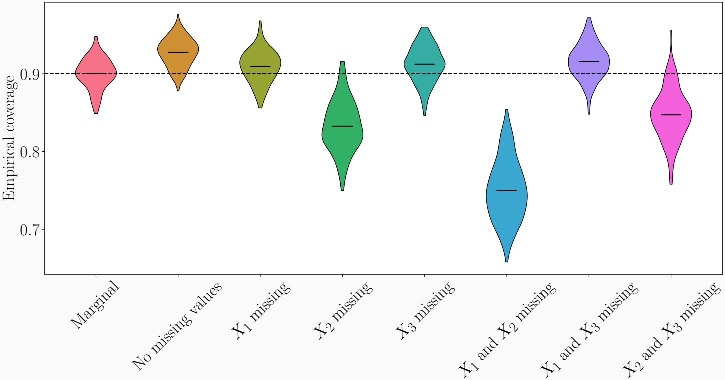
Even if the imputation is not accurate, the guarantee will hold.

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

CQR performances on an illustrative example

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

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



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

Missing values induce heteroskedasticity

Theoretical study of the Gaussian linear model ($Y = \beta^T X + \varepsilon$) generalizes:

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

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

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

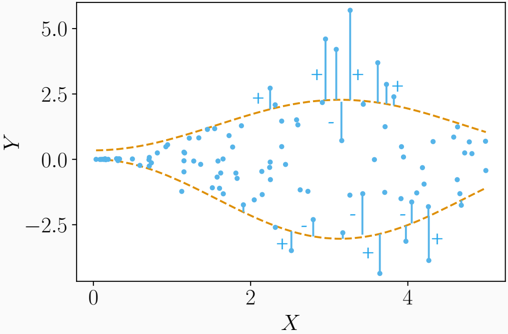
Goal: validity conditionally to the mask

Goal: for any $m \in \mathcal{M} \subset \{0,1\}^d$:

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

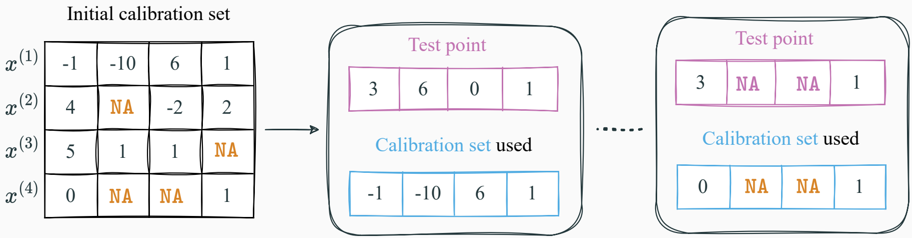
Motivation: equity, first-step-towards-conditional.

Issue during the calibration step



- Predict with $\widehat{QR}_{\alpha/2}$ and $\widehat{QR}_{1-\alpha/2}$
- Get the scores $\mathcal{S} = \left\{S^{(i)}\right\}_{\text{Cal}} \cup \{+\infty\}$
- Compute the $(1-\alpha)$ empirical quantile of $\mathcal{S}$, noted $q_{1-\alpha}\left(\mathcal{S}\right)$

Infeasible solution: splitting the calibration set[10] for each mask



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

Missing data augmentation of the calibration set

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

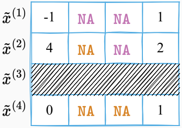
$\tilde{x}^{(1)}$	-1	NA	NA	1
$\tilde{x}^{(2)}$	4	NA	NA	2
$\tilde{x}^{(3)}$				
$\tilde{x}^{(4)}$	0	NA	NA	1

$$\hookrightarrow \quad S^{(i)} := \max\left\{\widehat{QR}_{\alpha/2}\left(\tilde{X}^{(i)}\right) - Y^{(i)}, Y^{(i)} - \widehat{QR}_{1-\alpha/2}\left(\tilde{X}^{(i)}\right)\right\}$$

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



   5.1 For each $j \in [\![1, d]\!]$ s.t. $M_j^{(n+1)} = 1$, set $\tilde{M}_j^{(i)} = 1$ for i in `Cal` s.t. $M^{(i)} \subset M^{(n+1)}$
   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, $q_{1-\alpha}(\mathcal{S})$

Mask conditional validity

Theorem (Zaffran et al., 2023)

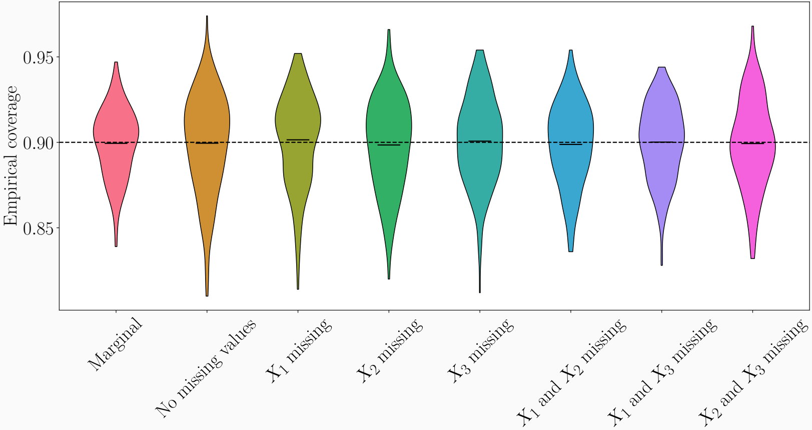
If the data is exchangeable and MCAR, then for almost all imputation function the proposed methodology is such that for any $m \in \{0,1\}^d$:

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

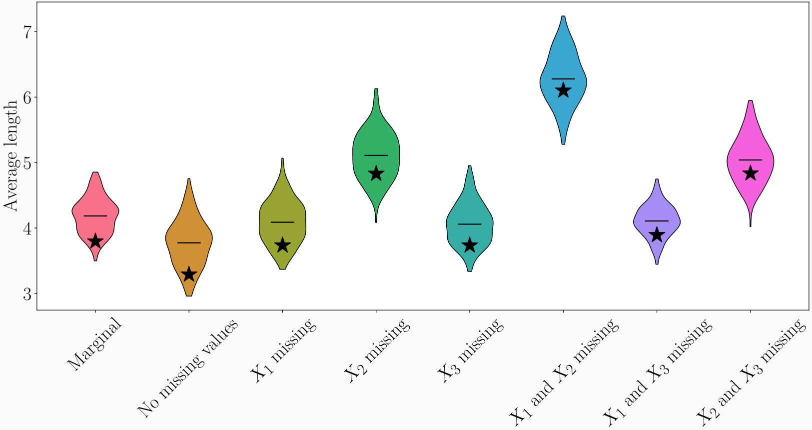
and if additionally the scores are almost surely distinct:

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

Empirical coverages



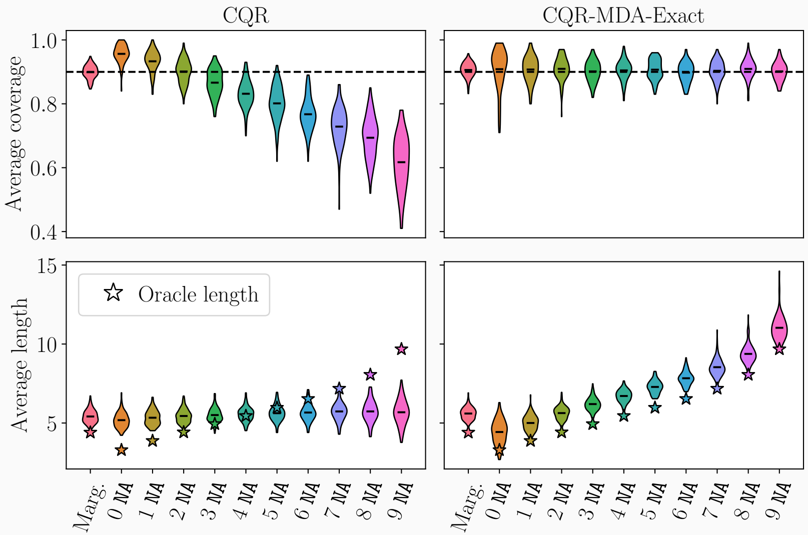
Empirical lengths



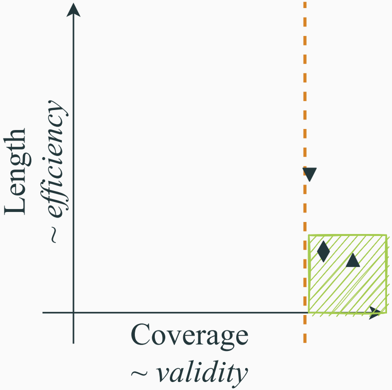
Some settings

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

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



Before more experiments, visualisation



♦ : marginal coverage, i.e.

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

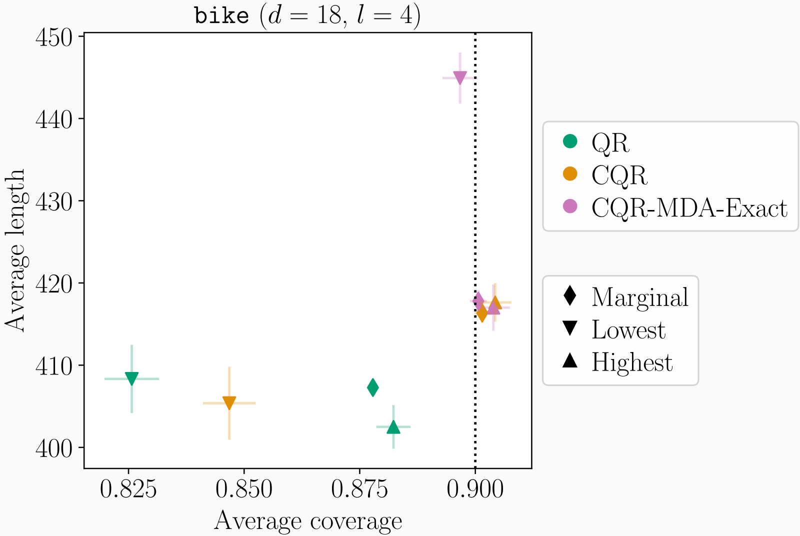
▼ : lowest coverage, i.e.

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

▲ : highest coverage, i.e.

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

Semi-synthetic experiments



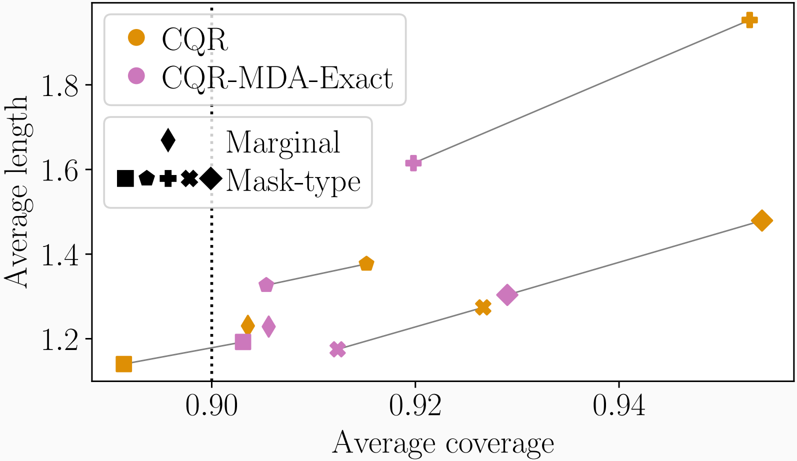
TraumaBase®: decision support for trauma patients

- 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



Extensions

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

Take-home-messages

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

Thank you! Questions? :)

Angelopoulos, A. N., Kohli, A. P., Bates, S., Jordan, M., Malik, J., Alshaabi, T., Upadhyayula, S., and Romano, Y. (2022). Image-to-image regression with distribution-free uncertainty quantification and applications in imaging. In *ICML*.

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

Barber, R. F., Candès, E. J., Ramdas, A., and Tibshirani, R. J. (2021a). The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*, 10(2).

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

Barber, R. F., Candès, E. J., Ramdas, A., and Tibshirani, R. J. (2023). Conformal prediction beyond exchangeability. *The Annals of Statistics*.

Bastani, O., Gupta, V., Jung, C., Noarov, G., Ramalingam, R., and Roth, A. (2022). Practical adversarial multivalid conformal prediction. In *NeurIPS*.

Candès, E., Lei, L., and Ren, Z. (2023). Conformalized survival analysis. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(1).

Cauchois, M., Gupta, S., Ali, A., and Duchi, J. C. (2020). Robust Validation: Confident Predictions Even When Distributions Shift. arXiv.

Chernozhukov, V., Wüthrich, K., and Yinchu, Z. (2018). Exact and Robust Conformal Inference Methods for Predictive Machine Learning with Dependent Data. In *COLT*.

Gibbs, I. and Candès, E. (2021). Adaptive conformal inference under distribution shift. In *NeurIPS*.

Gibbs, I. and Candès, E. (2022). Conformal inference for online prediction with arbitrary distribution shifts. arXiv.

Guan, L. (2022). Localized conformal prediction: a generalized inference framework for conformal prediction. *Biometrika*, 110(1).

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

Izbicki, R., Shimizu, G., and Stern, R. B. (2022). Cd-split and hpd-split: Efficient conformal regions in high dimensions. *Journal of Machine Learning Research*, 23(87).

Jin, Y., Ren, Z., and Candès, E. J. (2023). Sensitivity analysis of individual treatment effects: A robust conformal inference approach. *Proceedings of the National Academy of Sciences of the United States of America*, 120(6).

Jung, C., Noarov, G., Ramalingam, R., and Roth, A. (2023). Batch multivalid conformal prediction. In *ICLR*.

References v

Kath, C. and Ziel, F. (2021). Conformal prediction interval estimation and applications to day-ahead and intraday power markets. *International Journal of Forecasting*, 37(2).

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*.

Lei, J. and Wasserman, L. (2014). Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1).

Lei, L. and Candès, E. J. (2021). Conformal inference of counterfactuals and individual treatment effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83(5).

Lu, C., Angelopoulos, A. N., and Pomerantz, S. (2022). Improving trustworthiness of ai disease severity rating in medical imaging with ordinal conformal prediction sets. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*. Springer Nature Switzerland.

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

Podkopaev, A. and Ramdas, A. (2021). Distribution-free uncertainty quantification for classification under label shift. In *UAI*.

Romano, Y., Barber, R. F., Sabatti, C., and Candès, E. (2020). With Malice Toward None: Assessing Uncertainty via Equalized Coverage. *Harvard Data Science Review*, 2(2).

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

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

Schuster, T., Fisch, A., Gupta, J., Dehghani, M., Bahri, D., Tran, V. Q., Tay, Y., and Metzler, D. (2022). Confident adaptive language modeling. In *NeurIPS*.

Sesia, M. and Romano, Y. (2021). Conformal prediction using conditional histograms. In *NeurIPS*.

Taufiq, M. F., Ton, J.-F., Cornish, R., Teh, Y. W., and Doucet, A. (2022). Conformal off-policy prediction in contextual bandits. In *NeurIPS*.

Tibshirani, R. J., Barber, R. F., Candès, E., and Ramdas, A. (2019). Conformal prediction under covariate shift. In *NeurIPS*.

Vovk, V. (2012). Conditional Validity of Inductive Conformal Predictors. In *Asian Conference on Machine Learning*.

Vovk, V. (2015). Cross-conformal predictors. *Annals of Mathematics and Artificial Intelligence*, 74.

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

Zaffran, M., Féron, O., Goude, Y., Josse, J., and Dieuleveut, A. (2022). Adaptive conformal predictions for time series. In *ICML*.

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

Appendix

SCP in classification

SCP in classification

- $Y^{(i)} \in \{1, \ldots, C\}$ (C classes)
- $\hat{A}(X) = (\hat{p}_1(X), \ldots, \hat{p}_C(X))$ (estimated probabilities)
- Score of the i-th calibration point: $S^{(i)} = 1 - (\hat{A}\left(X^{(i)}\right))_{Y^{(i)}}$
- For a new point $X^{(n+1)}$, return

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

SCP in classification in practice

Ex: $Y^{(i)} \in \{$"dog", "tiger", "cat"$\}$, with $\alpha = 0.1$

- Scores on the calibration set

$\mathrm{Cal}^{(i)}$										
$\hat{p}_{\text{dog}}\left(X^{(i)}\right)$	0.95	0.90	0.85	0.15	0.15	0.20	0.15	0.15	0.25	0.20
$\hat{p}_{\text{tiger}}\left(X^{(i)}\right)$	0.02	0.05	0.10	0.60	0.55	0.50	0.45	0.40	0.35	0.45
$\hat{p}_{\text{cat}}\left(X^{(i)}\right)$	0.03	0.05	0.05	0.25	0.30	0.30	0.40	0.45	0.40	0.35
$s^{(i)}$	0.05	0.1	0.15	0.40	0.45	0.50	0.55	0.55	0.6	0.65

- $q_{1-\alpha}(\mathcal{S}) = 0.65$ $\lceil 0.9 \times (10+1) \rceil = 10$
- $\hat{A}\left(X^{(n+1)}\right) = (0.05, 0.60, 0.35)$
  - $\hookrightarrow$ $s(\hat{A}\left(X^{(n+1)}\right), \text{"dog"}) = 0.95$ "dog" $\notin \widehat{C}_\alpha\left(X^{(n+1)}\right)$
  - $\hookrightarrow$ $s(\hat{A}\left(X^{(n+1)}\right), \text{"tiger"}) = 0.40 \leq q_{1-\alpha}(\mathcal{S})$
    "tiger" $\in \widehat{C}_\alpha\left(X^{(n+1)}\right)$
  - $\hookrightarrow$ $s(\hat{A}\left(X^{(n+1)}\right), \text{"cat"}) = 0.65 \leq q_{1-\alpha}(\mathcal{S})$ "cat" $\in \widehat{C}_\alpha\left(X^{(n+1)}\right)$
- $\widehat{C}_\alpha\left(X^{(n+1)}\right) = \{$"tiger", "cat"$\}$

SCP in classification in practice

Ex: $Y^{(i)} \in \{$"dog", "tiger", "cat"$\}$, with $\alpha = 0.1$

- Scores on the calibration set

$\mathrm{Cal}^{(i)}$										
$\hat{p}_{\text{dog}}\left(X^{(i)}\right)$	0.95	0.90	0.85	0.05	0.05	0.05	0.05	0.10	0.10	0.15
$\hat{p}_{\text{tiger}}\left(X^{(i)}\right)$	0.02	0.05	0.10	0.85	0.80	0.75	0.70	0.25	0.30	0.30
$\hat{p}_{\text{cat}}\left(X^{(i)}\right)$	0.03	0.05	0.05	0.10	0.15	0.20	0.25	0.65	0.60	0.55
$S^{(i)}$	0.05	0.1	0.15	0.15	0.20	0.25	0.30	0.35	0.40	0.45

- $q_{1-\alpha}(\mathcal{S}) = 0.45$ $\quad \lceil 0.9 \times (10+1) \rceil = 10$
- $\hat{A}\left(X^{(n+1)}\right) = (0.05, 0.60, 0.35)$
  - $\hookrightarrow s(\hat{A}\left(X^{(n+1)}\right), \text{"dog"}) = 0.95$ $\quad$ "dog" $\notin \widehat{C}_\alpha\left(X^{(n+1)}\right)$
  - $\hookrightarrow s(\hat{A}\left(X^{(n+1)}\right), \text{"tiger"}) = 0.40 \leq q_{1-\alpha}(\mathcal{S})$
    "tiger" $\in \widehat{C}_\alpha\left(X^{(n+1)}\right)$
  - $\hookrightarrow s(\hat{A}\left(X^{(n+1)}\right), \text{"cat"}) = 0.65$ $\quad$ "cat" $\notin \widehat{C}_\alpha\left(X^{(n+1)}\right)$
- $\widehat{C}_\alpha\left(X^{(n+1)}\right) = \{$"tiger"$\}$

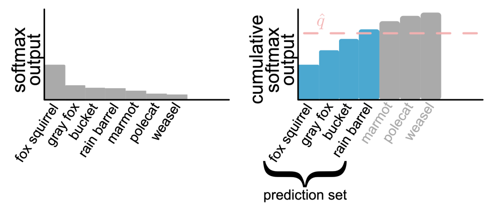
SCP in classification: comments on the naive version

- Facts about the previous method
  - prediction sets with the smallest average size
  - undercover hard subgroups
  - overcover easy ones
- Other types of scores can be used to improve the conditional coverage (as in regression with CQR or localized)

SCP in classification: Adaptive Prediction Sets

1. Sort in decreasing order $\hat{p}_{\sigma_i(1)}\left(X^{(i)}\right) \geq \ldots \geq \hat{p}_{\sigma_i(C)}\left(X^{(i)}\right)$
2. $S^{(i)} = \sum_{k=1}^{\sigma_i^{-1}\left(Y^{(i)}\right)} \hat{p}_{\sigma_i(k)}\left(X^{(i)}\right)$ (sum of the estimated probabilities associated to classes at least as large as that of the true class Y_i)
3. Return the classes $\sigma^{(n+1)}(1), \ldots, \sigma^{(n+1)}(r^\star)$ where

$$r^\star = \underset{1 \leq r \leq C}{\arg\max} \left\{ \sum_{k=1}^{r} \hat{p}_{\sigma^{(n+1)}(k)}\left(X^{(n+1)}\right) < q_{1-\alpha}(\mathcal{S}) \right\} + 1$$



SCP in classification in practice: Adaptive Prediction Sets

Ex: $Y_i \in \{\text{"dog"}, \text{"tiger"}, \text{"cat"}\}$, with $\alpha = 0.1$

- Scores on the calibration set

$\mathrm{Cal}^{(i)}$										
$\hat{p}_{\text{dog}}\left(X^{(i)}\right)$	0.95	0.90	0.85	0.05	0.05	0.05	0.10	0.25	0.10	0.15
$\hat{p}_{\text{tiger}}\left(X^{(i)}\right)$	0.02	0.05	0.10	0.85	0.80	0.75	0.75	0.40	0.30	0.30
$\hat{p}_{\text{cat}}\left(X^{(i)}\right)$	0.03	0.05	0.05	0.10	0.15	0.20	0.15	0.35	0.60	0.55
$S^{(i)}$	0.95	0.90	0.85	0.85	0.80	0.75	0.75	0.75	0.60	0.55

- $q_{1-\alpha}(\mathcal{S}) = 0.95$
- Ex 1: $\hat{A}\left(X^{(n+1)}\right) = (0.05, 0.45, 0.5)$, $r^\star = 2$

$$\widehat{C}_\alpha\left(X^{(n+1)}\right) = \{\text{"tiger"}, \text{"cat"}\}$$

- Ex 2: $\hat{A}\left(X^{(n+1)}\right) = (0.03, 0.95, 0.02)$, $r^\star = 1$

$$\widehat{C}_\alpha\left(X^{(n+1)}\right) = \{\text{"tiger"}\}$$

Jackknife/cross-val

Beyond the limitations of SCP

- SCP is computationally attractive: it only requires fitting the model one time
- Problem: it sacrifices statistical efficiency
  - requiring splitting the data into training and calibration datasets

$\rightsquigarrow$ Full (or transductive) conformal prediction
  - avoids data splitting
  - at the cost of many more model fits

- Historically, full conformal prediction was developed first
- Idea: we know that the true label $Y^{(n+1)}$ lives somewhere in $\mathcal{Y}$ so if we loop over all possible $y \in \mathcal{Y}$, then we will eventually hit the data point $(X^{(n+1)}, Y^{(n+1)})$, which is statistically plausible with the first n data points
- Hence the name as full conformal prediction directly computes this loop

Full conformal prediction

Method: for a candidate $(X^{(n+1)}, y)$,

1. Get $\hat{A}_y$ by training on $\{(X^{(1)}, Y^{(1)}), \dots, (X^{(n)}, Y^{(n)})\} \cup \{(X^{(n+1)}, y)\}$
2. Scores

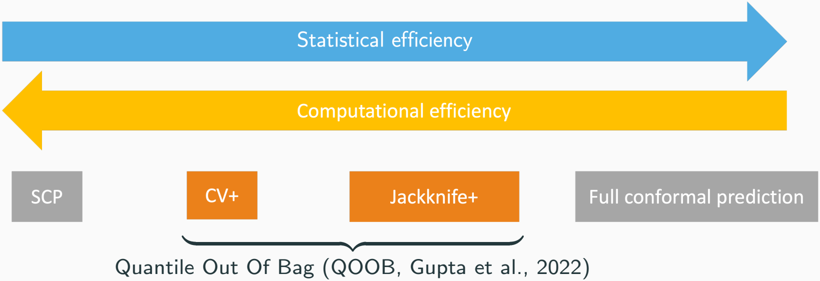
$$\mathcal{S} = \left\{ s(\hat{A}_y \left(X^{(i)}, Y^{(i)} \right) \right\} \cup \{ s(\hat{A}_y \left(X^{(n+1)} \right), y) \}$$

3. $y \in \widehat{C}_\alpha \left(X^{(n+1)} \right)$ if $s(\hat{A}_y \left(X^{(n+1)} \right), y) \leq q_{1-\alpha}(\mathcal{S})$

✓ Theoretical guarantees (provided that the learining algorithm handles exchangeable training data in a symmetric way)

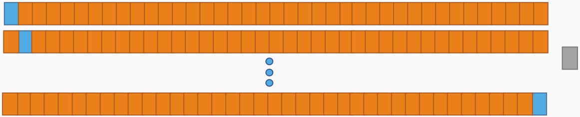
✗ Computationally costly: not used in practice

Other methods for conformal prediction



Jackknife: naive predictive interval

- Based on leave-one-out (LOO) residuals



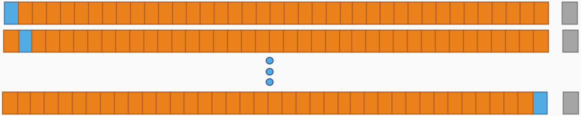
- $\mathcal{D}^n = \left\{ \left(X^{(1)}, Y^{(1)}\right), \ldots, \left(X^{(n)}, Y^{(n)}\right) \right\}$ training data
- Get $\hat{A}^{-i}$ by training on $\mathcal{D}^n \setminus \left(X^{(i)}, Y^{(i)}\right)$
- LOO scores $\mathcal{S} = \left\{ \left| \hat{A}^{-i}\left(X^{(i)}\right) - Y^{(i)} \right| \right\}_i \cup \{+\infty\}$ (in standard reg)
- Get $\hat{A}$ by training on $\mathcal{D}^n$
- Build the predictive interval: $\left[\hat{A}\left(X^{(n+1)}\right) \pm q_{1-\alpha}(\mathcal{S}) \right]$

Warning

No guarantee on the prediction of $\hat{A}$ with scores based on $(\hat{A}^{-i})_i$

Jackknife+ (Barber et al., 2021b)

- Based on leave-one-out (LOO) residuals



- $\mathcal{D}^n = \left\{ \left(X^{(1)}, Y^{(1)}\right), \ldots, \left(X^{(n)}, Y^{(n)}\right) \right\}$ training data
- Get $\hat{A}^{-i}$ by training on $\mathcal{D}^n \setminus \left(X^{(i)}, Y^{(i)}\right)$
- LOO predictions (in standard reg)
  $\mathcal{S}_{\text{up/down}} = \left\{ \hat{A}^{-i}\left(X^{(n+1)}\right) \pm \left|\hat{A}^{-i}\left(X^{(i)}\right) - Y^{(i)}\right| \right\}_i \cup \{\pm\infty\}$
- Build the predictive interval: $\left[q_{\alpha/2}(\mathcal{S}_{\text{down}}); q_{1-\alpha/2}(\mathcal{S}_{\text{up}})\right]$

Theorem

If $\mathcal{D}^n \cup (X^{(n+1)}, Y^{(n+1)})$ are exchangeable and the algorithm treats the data points symmetrically, then $\mathbb{P}(Y^{(n+1)} \in \widehat{C}_\alpha\left(X^{(n+1)}\right)) \geq 1 - 2\alpha$.

CV+ (Barber et al., 2021b)

- Based on cross-validation residuals
- $\mathcal{D}^n = \left\{ \left(X^{(1)}, Y^{(1)}\right), \ldots, \left(X^{(n)}, Y^{(n)}\right) \right\}$ training data

1. Split $\mathcal{D}^n$ into K folds $F_1, \ldots, F_K$
2. Get $\hat{A}^{-F_k}$ by training on $\mathcal{D}^n \setminus F_k$
3. Cross-val predictions (in standard reg)
   $$\mathcal{S}_{\text{up/down}} = \left\{ \left\{ \hat{A}^{-F_k}\left(X^{(n+1)}\right) \pm \left| \hat{A}_{-F_k}\left(X^{(i)}\right) - Y^{(i)} \right| \right\}_{i \in F_k} \right\}_k \cup \{\pm\infty\}$$
4. Build the predictive interval: $[q_\alpha(\mathcal{S}_{\text{down}}); q_{1-\alpha}(\mathcal{S}_{\text{up}})]$

Theorem

Under data exchangeability and algorithm symmetry, then
$\mathbb{P}(Y^{(n+1)} \in \widehat{C}_\alpha\left(X^{(n+1)}\right)) \geq 1 - 2\alpha - \min\left(\frac{2(1-1/K)}{n/K+1}, \frac{1-K/n}{K+1}\right) \geq 1 - 2\alpha - \sqrt{2/n}.$

CP-MDA-Nested

CP-MDA-Exact reminder

Test point

3	NA	NA	1

Initial calibration set

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

Calibration set used

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

What if we kept all individuals?

Test point

3	NA	NA	1

Initial calibration set

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

Calibration set used

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

Idea: modify the test point accordingly

Test point

3	NA	NA	1

Initial calibration set

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

Calibration set used

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

and

Temporary test points

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

CQR-MDA with nested masking in words

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

3	NA	NA	1

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

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

1.2 Impute the new calibration set

1.3 For each augmented calibration point i:

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

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

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

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

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

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

Summary of CP-MDA

Test point

3	NA	NA	1

Initial calibration set

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

CP-MDA with exact masking:

calibration set

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

CP-MDA with nested masking:

calibration set

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

and

temporary test points

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

Towards asymptotic individualized coverage

Consistency of a universal quantile learner after imputation

Let Φ be an imputation function chosen by the user.

Denote

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

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

Proposition (Pinball-consistency of an universal learner)

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

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

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

Asymptotic conditional coverage of a universal quantile learner

Corollary

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

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

$d = 3$

Data generation

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

$Y = \beta X + \varepsilon$

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

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

$Y = \beta X + \varepsilon$

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

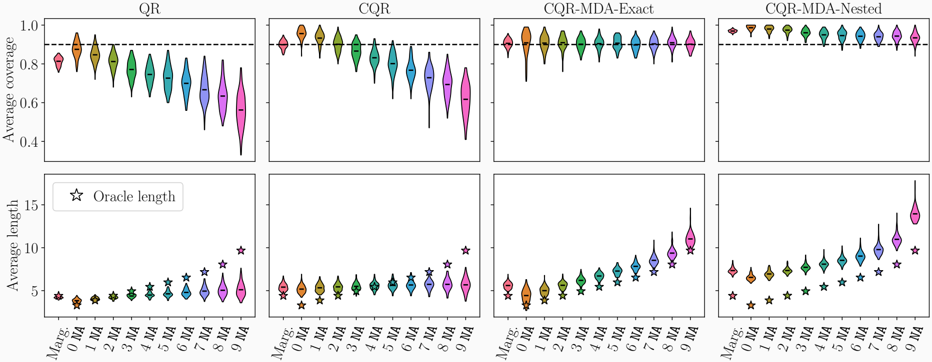
$$\text{and } (X_1, \cdots, X_{10}) \sim \mathcal{N}\left(\begin{pmatrix} 1 \\ \vdots \\ \vdots \\ 1 \end{pmatrix}, \begin{pmatrix} 1 & 0.8 & \cdots & 0.8 \\ 0.8 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0.8 \\ 0.8 & \cdots & 0.8 & 1 \end{pmatrix} \right).$$

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

Simulation settings

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

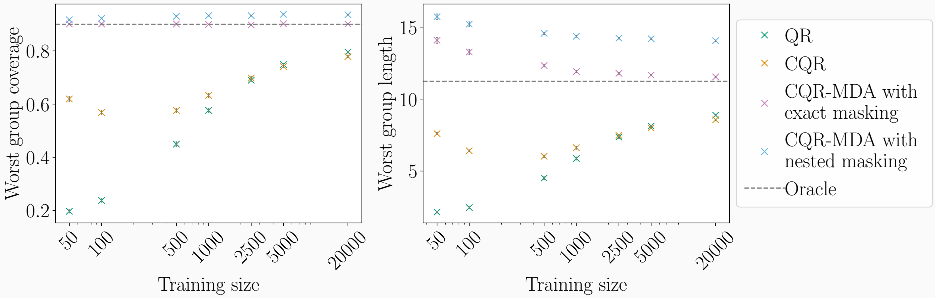
Results per pattern size



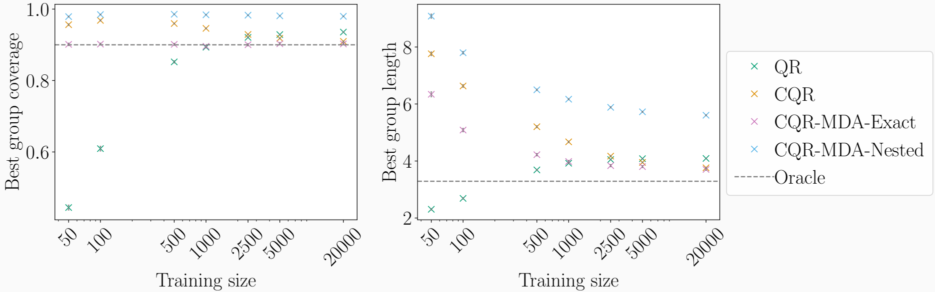
Simulation settings: varying training size

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

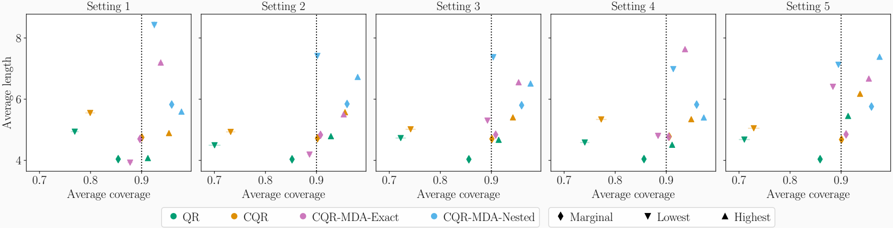
Results on the worst group



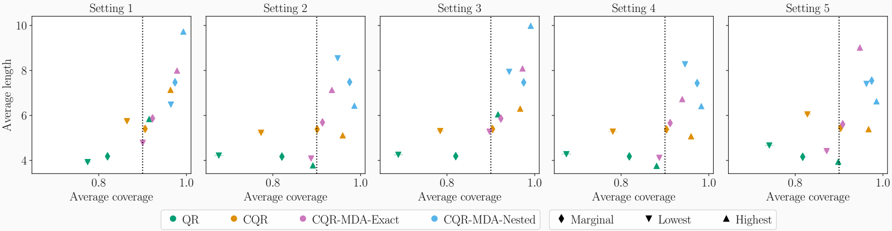
Results on the best group



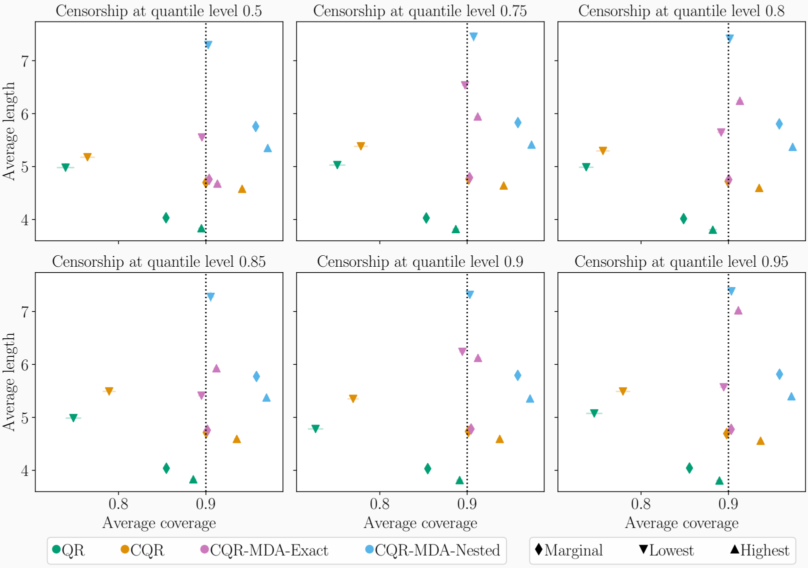
MAR missingness



MNAR self masked missingness

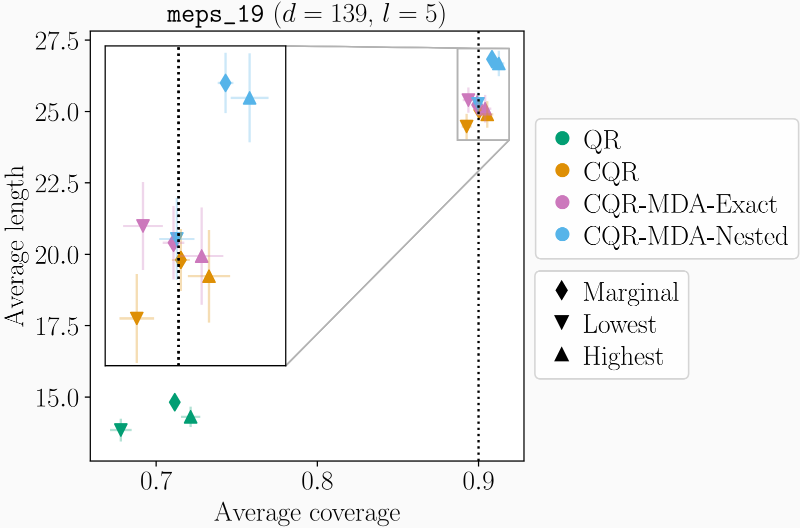


MNAR quantile censorship missingness

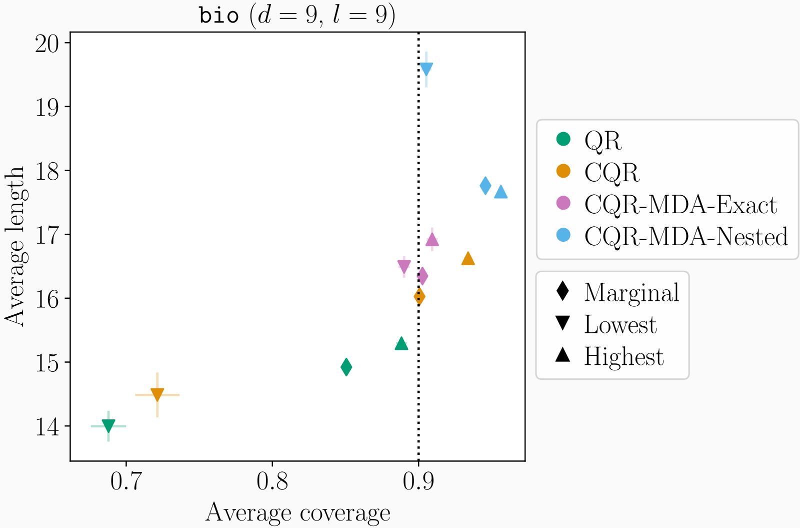


Semi-synthetic experiments with CQR-MDA-Nested

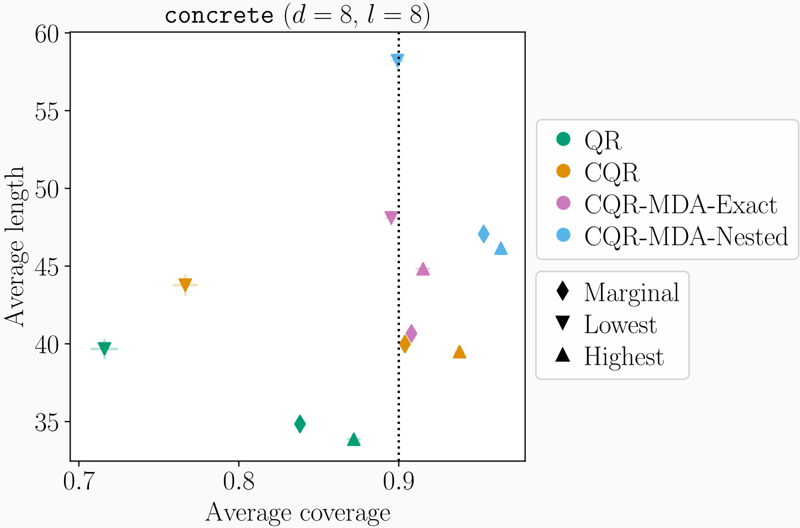
Semi-synthetic experiments with CQR-MDA-Nested



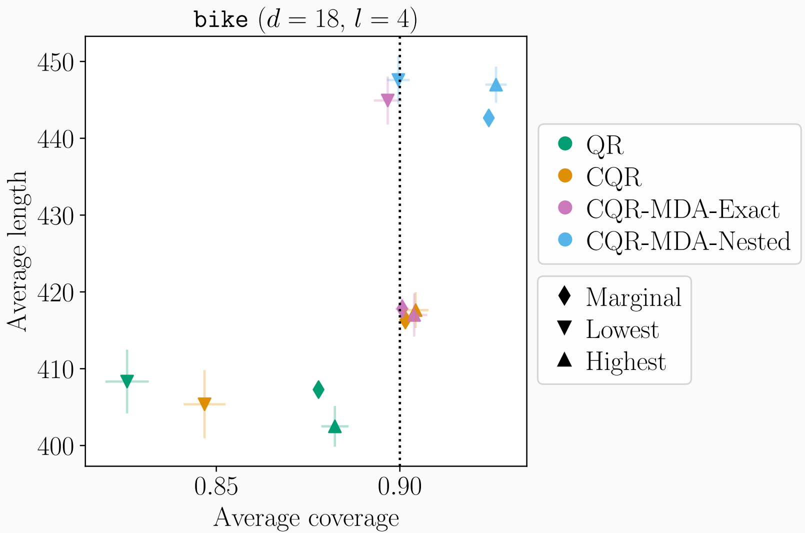
Semi-synthetic experiments with CQR-MDA-Nested



Semi-synthetic experiments with CQR-MDA-Nested



Semi-synthetic experiments with CQR-MDA-Nested



TraumaBase®

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.

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 $\text{SI} = \frac{\text{HR}}{\text{SBP}}$, upon arrival at hospital (2.09% missing values);
- `HR`: the heart rate measured upon arrival of hospital (1.62% missing values).

Results with CQR-MDA-Nested