Post-hoc predictive uncertainty quantification: methods with applications to electricity price forecasting

Margaux Zaffran — June 25, 2024 — Ph.D. defense

Ph.D. advisors

Aymeric Dieuleveut
Julie Josse
Olivier Féron
Yannig Goude

Jury members

Reviewers

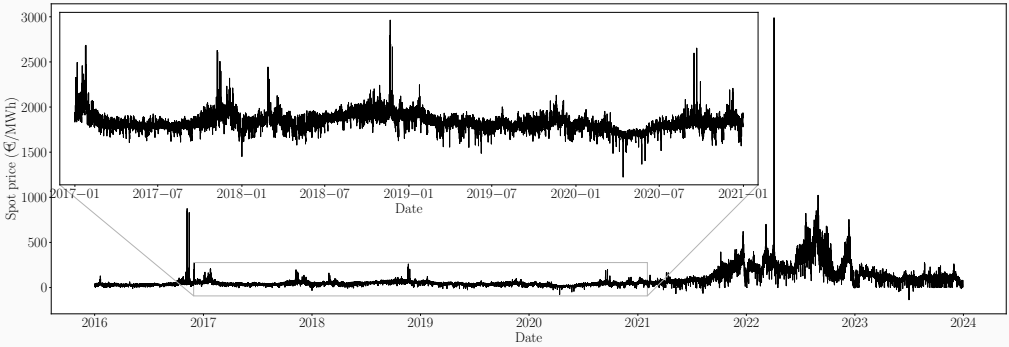
Pierre Pinson
Étienne Roquain

Examiners

Emmanuel Candès
Florence Forbes
Éric Moulines
Aaditya Ramdas

Forecasting French spot electricity prices

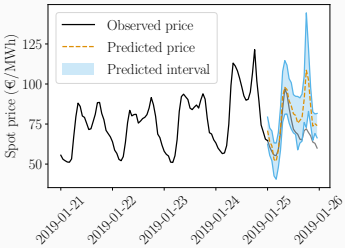
Hourly day-ahead market prices (between producers and suppliers)



To which extent are they forecastable?

↪ forecasts errors **no lower than 10%** of the realized price!

New goal:



Quantify predictive uncertainty with:

- **Theoretically grounded tools**
- **Light assumptions on the underlying data distribution**
- **Guarantees agnostic to the prediction algorithm**
  $\rightsquigarrow$ Post-hoc approach (i.e. no modification of the existing operational pipeline)

Time series

- ▷ Temporal structure (trend, seasonality, dependence, etc.)
- ▷ Non-stationarity

Missing values

Improve forecasts by leveraging the *emergence of open data platforms* (ENTSO-E Transparency, Eco2Mix, etc.)

- ▷ Missing covariates by aggregating different data sources

Approach: black-box post-processing of existing probabilistic forecasts

Important literature on intervals forecast, emerging from the electrical application (Hong et al., 2016; Hong and Fan, 2016), but also from renewable energies and meteorology (Wan et al., 2014; Wang et al., 2017).

Wide range of models, mainly based on the *pinball loss*, such as

- Quantile Random Forests,
- Quantile Generalized Additive Models,
- Quantile Regression Averaging,
- intervals from Gaussian Auto-Regressive models with exogenous variables,
- Deep Learning Probabilistic,

etc. $\rightsquigarrow$ in practice uncalibrated.

Black-box post-processing of available probabilistic forecasts

- Post-hoc approach: plug-in on top of any of these models
- Guarantees: in finite sample and distribution-free

Quantifying predictive uncertainty

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

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

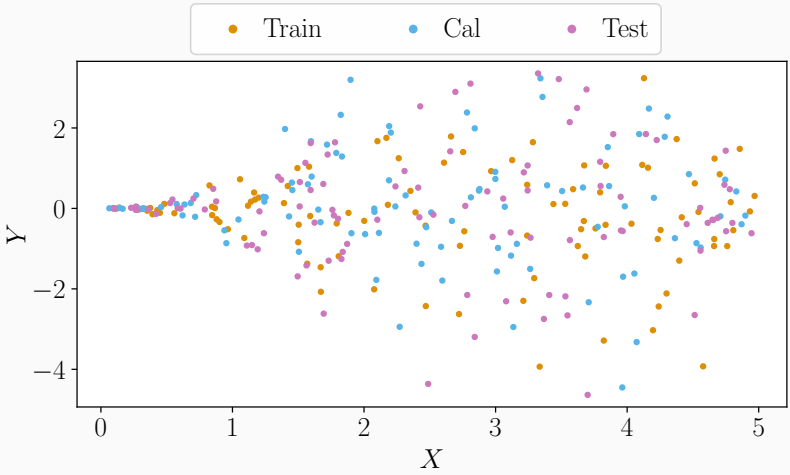
  and $\mathcal{C}_\alpha$ should be as small as possible, in order to be informative[1].

- Construction of the predictive intervals should be
  - agnostic to the model[2]
  - agnostic to the data distribution
  - valid in finite samples

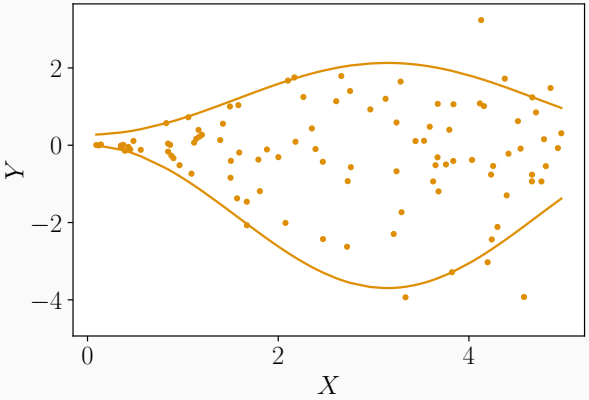
[1] Analogous to Gneiting et al. (2007).

[2] The underlying model can be any probabilistic model tailored for the application task at hand.

Conformalized Quantile Regression (CQR)[3]

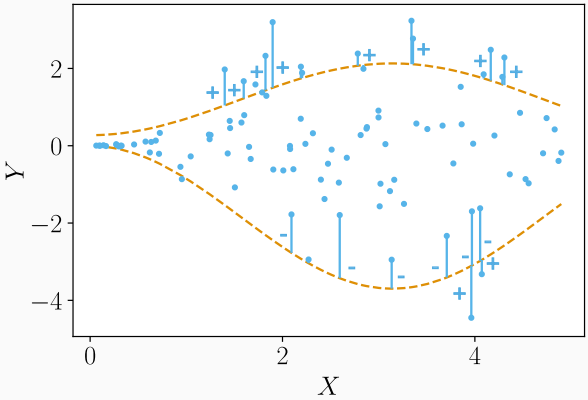


[3]Romano et al. (2019), *Conformalized Quantile Regression*, NeurIPS



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

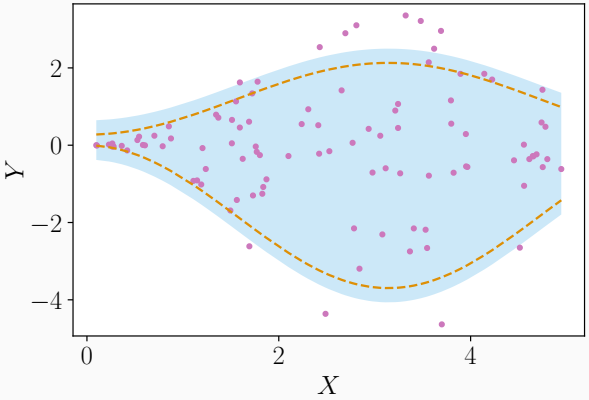
[3] Romano et al. (2019), *Conformalized Quantile Regression*, NeurIPS



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

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

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



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

- Build

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

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

Exchangeability

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

$$\left(X^{(1)}, Y^{(1)}\right), \ldots, \left(X^{(n)}, Y^{(n)}\right) \overset{d}{=} \left(X^{(\sigma(1))}, Y^{(\sigma(1))}\right), \ldots, \left(X^{(\sigma(n))}, Y^{(\sigma(n))}\right).$$

$\hookrightarrow$ i.i.d. $\Rightarrow$ exchangeability

CQR marginal validity (Romano et al., 2019)

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

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

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

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

Definition of distribution-free features conditional validity

$\widehat{C}_\alpha$ = estimated predictive set based on n data points.

Distribution-free X-conditional validity

$\widehat{C}_\alpha$ achieves distribution-free X-conditional validity if:

- for any distribution $\mathcal{D}$,
- for any associated exchangeable joint distribution $\mathcal{D}^{\text{exch}(n+1)}$,

we have that:

$$\mathbb{P}_{\mathcal{D}^{\text{exch}(n+1)}}\left(Y^{(n+1)} \in \widehat{C}_\alpha\left(X^{(n+1)}\right) | X^{(n+1)}\right) \overset{a.s.}{\geq} 1-\alpha.$$

Informative conditional coverage as such is impossible

Impossibility results (Vovk, 2012; Lei and Wasserman, 2014)

If $\widehat{C}_\alpha$ is distribution-free X-conditionally valid, then, for any $\mathcal{D}$, for $\mathcal{D}_X$–almost all $\mathcal{D}_X$–non-atoms $x \in \mathcal{X}$, it holds:

$$\mathbb{P}_{\mathcal{D}^{\otimes(n)}} \left\{ \text{mes}\left(\widehat{C}_\alpha(x) \right) = \infty \right\} \geq 1 - \alpha.$$

- Approximate conditional coverage
  $\hookrightarrow$ Romano et al. (2020); Guan (2022); Jung et al. (2023); Gibbs et al. (2023)
  Target $\mathbb{P}(Y^{(n+1)} \in \widehat{C}_\alpha\left(X^{(n+1)}\right) | X^{(n+1)} \in \mathcal{R}(x)) \geq 1 - \alpha$
- Asymptotic (with the sample size) conditional coverage
  $\hookrightarrow$ **Romano et al. (2019)**; Kivaranovic et al. (2020); Chernozhukov et al. (2021); Sesia and Romano (2021); Izbicki et al. (2022)

Non exhaustive references.

Contributions and outline

Part I: time series

▷ Adaptive Conformal Predictions for Time Series. In *ICML*.
(Z., Féron, Goude, Josse, and Dieuleveut, 2022)

▷ Adaptive Probabilistic Forecasting of French Electricity Spot Prices. Submitted to *Applied Energy*. (Dutot*, Z.*, Féron, and Goude, 2024)

Part II: missing values

▷ Conformal Prediction with Missing Values. In *ICML*.
(Z., Dieuleveut, Josse, and Romano, 2023)

▷ Predictive Uncertainty Quantification with Missing Covariates. Submitted to *Journal of Machine Learning Research*. (Z., Josse, Romano, and Dieuleveut, 2024)

Online framework

- Data: T_0 random variables $(X^{(1)}, Y^{(1)}), \ldots, (X^{(T_0)}, Y^{(T_0)})$ in $\mathbb{R}^d \times \mathbb{R}$
- Aim: predict the response values as well as predictive intervals for T_1 subsequent observations $X^{(T_0+1)}, \ldots, X^{(T_0+T_1)}$ **sequentially**:
  at any prediction step $t \in [\![T_0 + 1, T_0 + T_1]\!]$, $Y^{(t-T_0)}, \ldots, Y^{(t-1)}$ have been revealed
- Build the smallest interval $\widehat{C}_\alpha^t$ such that:

$$\mathbb{P}\left\{Y^{(t)} \in \widehat{C}_\alpha^t\left(X^{(t)}\right)\right\} \geq 1 - \alpha, \quad \text{for } t \in [\![T_0 + 1, T_0 + T_1]\!],$$

  often relaxed in:

$$\frac{1}{T_1} \sum_{t=T_0+1}^{T_0+T_1} \mathbb{1}\left\{Y^{(t)} \in \widehat{C}_\alpha^t\left(X^{(t)}\right)\right\} \approx 1 - \alpha.$$

Extensions of CP to forecasting time series (as of 2021)

- Theory (Chernozhukov et al., 2018)
- Applications (Wisniewski et al., 2020; Kath and Ziel, 2021)
- Gibbs and Candès (2021)

- Theory (Chernozhukov et al., 2018)
- Applications (Wisniewski et al., 2020; Kath and Ziel, 2021)
- Gibbs and Candès (2021)

Adaptive Conformal Inference (ACI) was initially proposed to handle distribution shift.

It relies on updating online an *effective miscoverage rate* α_t, with the scheme

$$\alpha_{t+1} := \alpha_t + \gamma \left(\alpha - \mathbb{1} \left\{ Y^{(t)} \notin \widehat{C}_{\alpha_t} \left(X^{(t)} \right) \right\} \right),$$

and $\alpha_1 = \alpha$, $\gamma \geq 0$.

Intuition: if we did make an error, the interval was too small so we want to increase its length by taking a higher quantile (a smaller α_t). Reversely if we included the point.

Visualisation of ACI procedure

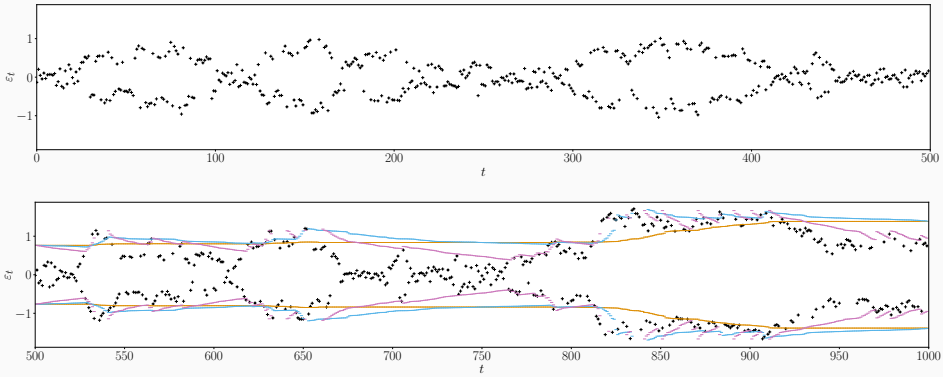


Figure 1: Visualisation of ACI with different values of γ ($\gamma = 0$, $\gamma = 0.01$, $\gamma = 0.05$)

Gibbs and Candès (2021) provide an asymptotic validity result for any sequence of observations.

$$\left| \frac{1}{T_1} \sum_{t=T_0+1}^{T_0+T_1} \mathbb{1}\left\{ Y^{(t)} \in \widehat{C}_{\alpha_t}\left(X^{(t)}\right) \right\} - (1-\alpha) \right| \leq \frac{2}{\gamma T_1}$$

$\Rightarrow$ favors large γ. But, the higher γ, the more frequent are the infinite intervals.

Approach

Aim: derive theoretical results on the average length of ACI depending on γ

$\hookrightarrow$ guideline for choosing γ

Approach:

- consider extreme cases (useful in an online context) with simple theoretical distributions
  1. exchangeable
  2. Auto-Regressive case (AR(1))
- assume the calibration is perfect, to rely on Markov Chain Theory

  $\hookrightarrow$ the empirical quantiles correspond to the exact scores' quantile distribution Q

Theoretical analysis of ACI's length: exchangeable case

Define:

- $L(\alpha_t) = 2Q(1-\alpha_t)$ the adaptive algorithm's interval's length at time t,
- $L_0 = 2Q(1-\alpha)$ the non-adaptive algorithm's interval's length (i.e. $\gamma = 0$).

Limit length under exchangeability (Z., Féron, Goude, Josse, and Dieuleveut, 2022)

Assume the scores are exchangeable with quantile function Q perfectly estimated at each time, and other technical assumptions.
Then, for all $\gamma > 0$, $(\alpha_t)_{t>0}$ forms a Markov Chain, that admits a stationary distribution π_γ, and

$$\frac{1}{T}\sum_{t=1}^{T} L(\alpha_t) \underset{T\to+\infty}{\xrightarrow{a.s.}} \mathbb{E}_{\pi_\gamma}[L] \overset{\text{not.}}{=} \mathbb{E}_{\tilde{\alpha}\sim\pi_\gamma}[L(\tilde{\alpha})].$$

Moreover, as $\gamma \to 0$, $\mathbb{E}_{\pi_\gamma}[L] = L_0 + Q''(1-\alpha)\frac{\gamma}{2}\alpha(1-\alpha) + O(\gamma^{3/2})$.

Convergence under AR(1) (Z., Féron, Goude, Josse, and Dieuleveut, 2022)

Assume the residuals follow an AR(1) process: $\hat{\varepsilon}^{(t+1)} = \varphi \hat{\varepsilon}^{(t)} + \xi^{(t+1)}$ with $(\xi^{(t)})_t$ i.i.d. random variables and other technical assumptions, we have:

$$\frac{1}{T}\sum_{t=1}^{T} L(\alpha_t) \xrightarrow[T\to+\infty]{a.s.} \mathbb{E}_{\pi_{\gamma,\varphi}}[L] \overset{\text{not.}}{=} \mathbb{E}_{\tilde{\alpha}\sim\pi_{\gamma,\varphi}}[L(\tilde{\alpha})].$$

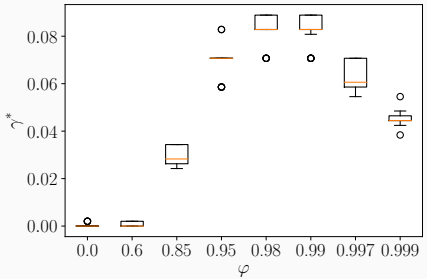


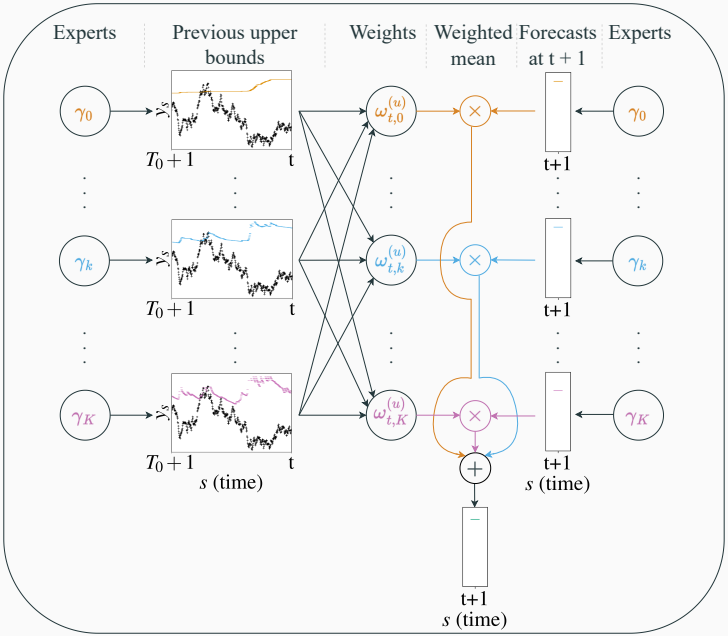
Figure 2: γ^* minimizing the average length for each φ.

AgACI: adaptive wrapper around ACI

Online aggregation under expert advice (Cesa-Bianchi and Lugosi, 2006) computes an optimal weighted mean of experts.

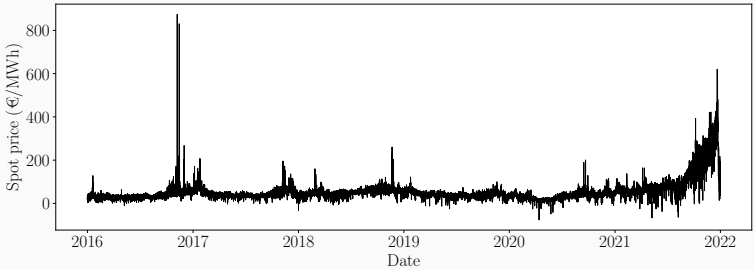
AgACI performs 2 independent aggregations: one for each bound (the upper and lower ones), based on the corresponding pinball losses.

AgACI: adaptive wrapper around ACI, scheme (upper bound)



Experimental take-away messages

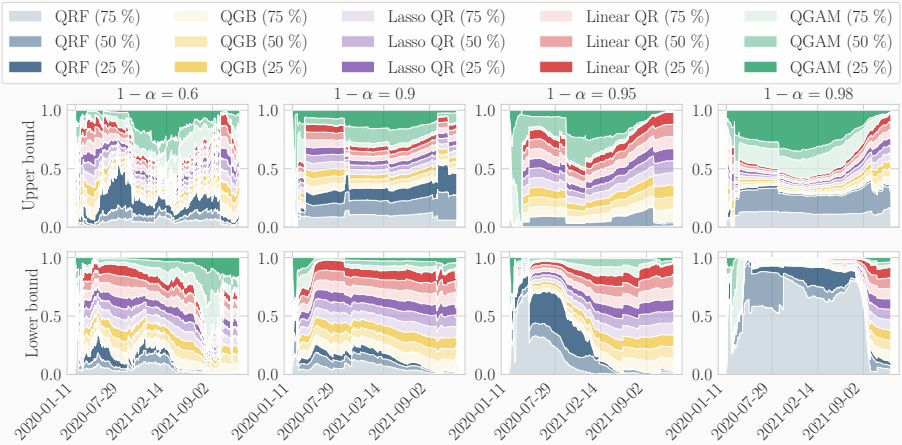
- **Synthetic data with ARMA noise** (Z., Féron, Goude, Josse, and Dieuleveut, 2022)
  - Benchmarks are not robust to the increase in the temporal dependence;
  - ACI is robust, maintaining validity, with an appropriate γ;
  - `AgACI` is robust, maintaining validity, not the smallest.
- **French electricity spot prices**
  - 2019: `AgACI` provides validity with a reasonable efficiency; (Z., Féron, Goude, Josse, and Dieuleveut, 2022)
  - 2020 and 2021: `AgACI` fails to ensure validity, and the various forecasting models considered behave differently. (Dutot*, Z.*, Féron, and Goude, 2024)



Improving adaptiveness for high non-stationarity (Dutot*, Z.*, Féron, and Goude, 2024)

Online aggregation of various `AgACI`, each of them being trained with different underlying forecasting models, for each bound independently.

- ✓ Retrieves validity even in the most hazardous period of 2020 and 2021.
- ✓ Analyzing its weights provides interpretability.



Aggregating the two bounds independently (as in `AgACI` and beyond):

- ✓ Allows more flexible and adaptive behavior in practice, catching the varying nature of the predictive distribution tails
- ✗ Prevents from obtaining theoretical guarantees (by opposition to Gibbs and Candès, 2022)

↪ Weaken the objective and consider a more practical theoretical aim?

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	Mask $M =$ (M_1	M_2	M_3)
22.42	0.55	0.67	0.03	0	0	0
8.26	0.72	0.18	0.55	0	0	0
19.41	0.60	0.58	NA	0	0	1
19.75	0.54	0.43	0.96	0	0	0
7.32	NA	0.19	NA	1	0	1
13.55	0.65	0.69	0.50	0	0	0
20.75	NA	NA	0.61	1	1	0
9.26	0.89	NA	0.84	0	1	0
9.68	0.963	0.45	0.65	0	0	0

$\hookrightarrow 2^d$ potential masks.

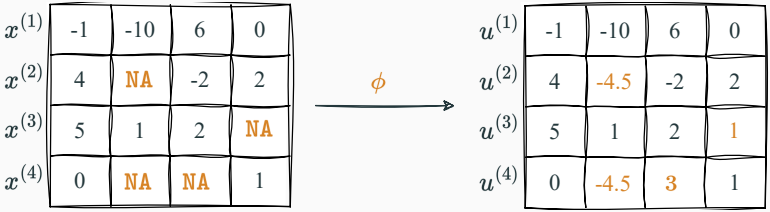
$\hookrightarrow M$ can depend on X or Y.

$\Rightarrow$ Statistical and computational challenges.

Supervised learning with missing values: impute-then-regress

Impute-then-regress procedures are widely used.

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



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

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

Goals of 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 for any $\mathcal{D}$ and any associated $\mathcal{D}^{\text{exch}(n+1)}$:

Marginal Validity (MV)

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

Mask-Conditional-Validity (MCV)

$$\mathbb{P}_{\mathcal{D}^{\text{exch}(n+1)}}\left\{Y^{(n+1)} \in \mathcal{C}_\alpha\left(X^{(n+1)}, M^{(n+1)}\right) | M^{(n+1)}\right\} \overset{a.s.}{\geq} 1-\alpha. \qquad \text{(MCV)}$$

M forms **meaningful categories**

$\hookrightarrow$ Even if $M \perp\!\!\!\perp X$ and $Y \perp\!\!\!\perp M \,|X$ (e.g. **equity standpoint**)

CP is marginally valid (MV) after imputation

Exchangeability after imputation (Z., Dieuleveut, Josse and Romano, 2023)

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

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

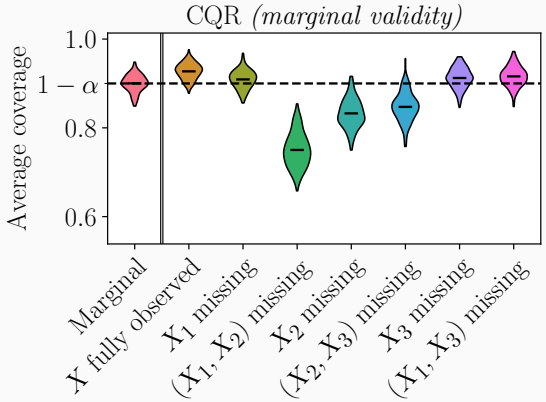
$\Rightarrow$ CQR, and Conformal Prediction, applied on an imputed data set still enjoys marginal guarantees[4]:

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

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

CQR is marginally valid on imputed data sets

$Y = \beta^T X + \varepsilon$, $\beta = (1, 2, -1)^T$, X and ε Gaussian.



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

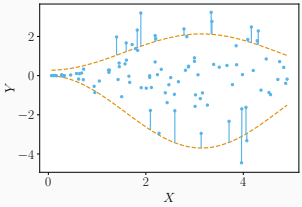
✗ Mask-conditional-validity (MCV) is not attained

↪ Missing values induce heteroskedasticity

(supported by theory under (non-)parametric assumptions)

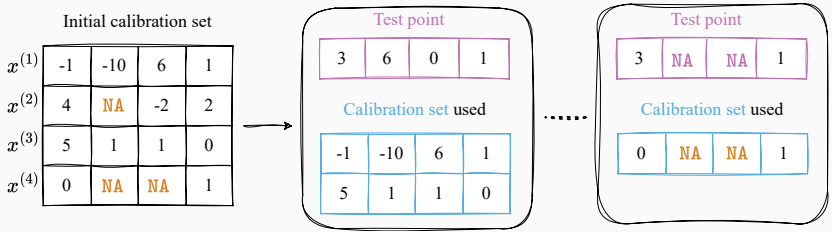
Conformalization step is independent of the important variable: the mask!

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 is infeasible (lack of data)!



Question: for low probability masks (i.e. $\mathcal{D}_M(m) := \mathbb{P}_{\mathcal{D}}(M = m)$ is small), is it possible to learn from the predictive distributions conditional on other masks?

Fully distribution-free MCV is necessarily uninformative

General MCV hardness result (Z., Josse, Romano and Dieuleveut, 2024)[5]

If any $\widehat{C}_\alpha$ is distribution-free MCV then **for any distribution** $\mathcal{D}$, for any mask m such that $\mathcal{D}_M(m) := \mathbb{P}_{\mathcal{D}}(M = m) > 0$, it holds:

$$\mathbb{P}_{\mathcal{D}^{\otimes(n+1)}}\left(\text{mes}\left(\widehat{C}_\alpha\left(X^{(n+1)}, m\right)\right) = \infty\right) \geq 1 - \alpha - \Delta_{m,n} \geq 1 - \alpha - \mathcal{D}_M(m)\sqrt{n+1}.$$

Irreducible term: consider $\widehat{C}_\alpha$ outputting $\mathcal{Y}$ with probability $1 - \alpha$ and $\emptyset$ otherwise.

$\Delta_{m,n}$ term: smaller than $\mathcal{D}_M(m)\sqrt{n+1}$

$\hookrightarrow$ gets negligible (making the lower bound nearly $1 - \alpha$) **only** for low probability masks compared to n.

[5] An analogous statement is also available for the classification framework.

Restricting the link between M and (X or Y) does not allow informative MCV

$M \perp\!\!\!\perp X$ hardness result (Z., Josse, Romano and Dieuleveut, 2024)

If any $\widehat{C}_\alpha$ is MCV under $M \perp\!\!\!\perp X$, then for any distribution $\mathcal{D}$ such that $M \perp\!\!\!\perp X$, for any mask m such that $\mathcal{D}_M(m) > 0$, it holds:

$$\mathbb{P}_{\mathcal{D}^{\otimes(n+1)}}\left(\text{mes}\left(\widehat{C}_\alpha\left(X^{(n+1)}, m\right)\right) = \infty\right) \geq 1 - \alpha - \mathcal{D}_M(m)\sqrt{n+1}.$$

$Y \perp\!\!\!\perp M \,|X$ hardness result (Z., Josse, Romano and Dieuleveut, 2024)

If any $\widehat{C}_\alpha$ is MCV under $Y \perp\!\!\!\perp M \,|X$, then for any distribution $\mathcal{D}$ such that $Y \perp\!\!\!\perp M \,|X$, for any mask m such that $\frac{1}{\sqrt{2}} \geq \mathcal{D}_M(m) > 0$, it holds:

$$\mathbb{P}_{\mathcal{D}^{\otimes(n+1)}}\left(\text{mes}\left(\widehat{C}_\alpha\left(X^{(n+1)}, m\right)\right) = \infty\right) \geq 1 - \alpha - 2\mathcal{D}_M(m)\sqrt{n+1}.$$

$\Rightarrow$ Need to restrict **both** the link between M and X, **as well as** between M and Y.

Analogous statements are also available for the classification framework.

CP-MDA-Nested* (Missing Data Augmentation): three instances

Mask-conditional-validity of CP-MDA-Nested*
(Z., Josse, Romano and Dieuleveut, 2024)

Under the assumptions that:

- $M \perp\!\!\!\perp X$,
- $Y \perp\!\!\!\perp M \,|X$,
- $\left(X^{(k)}, M^{(k)}, Y^{(k)}\right)_{k=1}^{n+1}$ are i.i.d.,
- the subsampling scheme is independent of $\left(X^{(k)}, Y^{(k)}\right)_{k=1}^{n+1}$,

then, for almost all imputation function, CP-MDA-Nested* reaches (MCV) at the level $1-2\alpha$, that is:

$$\mathbb{P}_{\mathcal{D}^{\otimes(n+1)}}\left\{Y^{(n+1)} \in \widehat{C}_\alpha\left(X^{(n+1)}, M^{(n+1)}\right) | M^{(n+1)}\right\} \overset{a.s.}{\geq} 1-2\alpha.$$

Experiments on $M \perp\!\!\!\perp X$ and $Y \perp\!\!\!\perp M \,|X$ Gaussian linear data in dimension 10

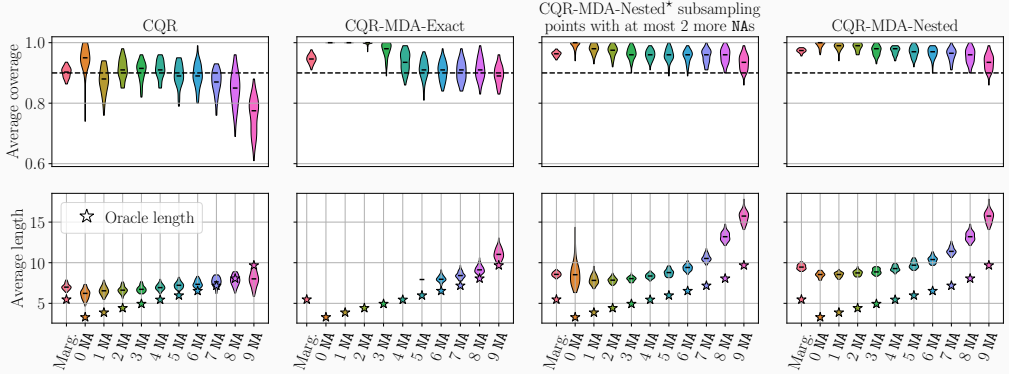


Figure 4: 40% of missing values

$\rightsquigarrow$ `CP-MDA-Exact` outputs many **infinite intervals** on points with less than 6 `NAs`.

$\rightsquigarrow$ Compared to `CP-MDA-Nested`, `CP-MDA-Nested`* selecting points with at most 2 more `NAs` **reduces the length** by:

- 5.5% marginally;
- 10% on fully observed points.

Experiments beyond independence

- ✓ Under various $M \not\perp\!\!\!\perp X$ (MAR and MNAR) mechanisms, `CP-MDA-Nested`* maintains empirical MCV;
- ✗ When $Y \not\perp\!\!\!\perp M \,|X$ and the imputation is not accurate enough:
  - `CP-MDA-Nested`* fails to empirically ensure MCV,
  - with a loss of coverage that is more critical when subsampling.

Key messages and contributions

Part I: time series

- ▷ Impact of hyper-parameter on the intervals efficiency
- ▷ Methodologies for online forecasting with post-hoc predictive UQ
- ▷ Extensive benchmark on time series CP and French elec. spot prices

Part II: missing values

- ▷ Missingness and predictive uncertainty interplay's characterization
- ▷ Methodology to achieve MCV
- ▷ Numerical experiments beyond the assumptions

Open-sourced introductive tutorial on CP, *(to be)* presented at:

- ▷ MASPIN days 2023, with C. Boyer,
- ▷ ENBIS 2023,
- ▷ UAI 2024, with A. Dieuleveut,
- ▷ ICML 2024, with A. Dieuleveut.

Some direct open directions include:

▷ Deeper investigation of practical time series CP (data sets, extremes, improved model, interpertability, theoretical objective)

▷ Is it possible to informatively achieve MCV under Missing At Random and $Y \perp\!\!\!\perp M \,|X$?

▷ Multidimensional predictive uncertainty quantification

Motivation: forecast multiple (correlated) electricity prices simultaneously (e.g., different countries or market horizons)

Challenge: capture the multivariate uncertainty

Thank you for your attention!

And many thanks to

Aymeric Dieuleveut

Olivier Féron

Yannig Goude

Julie Josse

Claire Boyer

Grégoire Dutot

Yaniv Romano

and many others :)

Angelopoulos, A. N., Candès, E. J., and Tibshirani, R. J. (2023). Conformal pid control for time series prediction. arXiv: 2307.16895.

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

Barber, R. F., Candès, E. J., Ramdas, A., and Tibshirani, R. J. (2022). Conformal prediction beyond exchangeability. To appear in *Annals of Statistics (2023)*.

Bastani, O., Gupta, V., Jung, C., Noarov, G., Ramalingam, R., and Roth, A. (2022). Practical adversarial multivalid conformal prediction. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc.

Bhatnagar, A., Wang, H., Xiong, C., and Bai, Y. (2023). Improved online conformal prediction via strongly adaptive online learning. In *Proceedings of the 40th International Conference on Machine Learning*. PMLR.

References ii

Cauchois, M., Gupta, S., Ali, A., and Duchi, J. C. (2020). Robust Validation: Confident Predictions Even When Distributions Shift. arXiv: 2008.04267.

Cesa-Bianchi, N. and Lugosi, G. (2006). *Prediction, learning, and games*. Cambridge University Press.

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

Chernozhukov, V., Wüthrich, K., and Zhu, Y. (2021). Distributional conformal prediction. *Proceedings of the National Academy of Sciences*, 118(48).

Feldman, S., Ringel, L., Bates, S., and Romano, Y. (2023). Achieving risk control in online learning settings. *Transactions on Machine Learning Research (TMLR)*.

Gibbs, I. and Candès, E. (2021). Adaptive conformal inference under distribution shift. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc.

Gibbs, I. and Candès, E. (2022). Conformal inference for online prediction with arbitrary distribution shifts. arXiv: 2208.08401.

Gibbs, I., Cherian, J. J., and Candès, E. J. (2023). Conformal prediction with conditional guarantees. arXiv: 2305.12616.

Gneiting, T., Balabdaoui, F., and Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 69(2):243–268.

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

Hong, T. and Fan, S. (2016). Probabilistic electric load forecasting: A tutorial review. *International Journal of Forecasting*, 32(3):914–938.

Hong, T., Pinson, P., Fan, S., Zareipour, H., Troccoli, A., and Hyndman, R. J. (2016). Probabilistic energy forecasting: Global energy forecasting competition 2014 and beyond. *International Journal of Forecasting*, 32(3):896–913.

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).

Jung, C., Noarov, G., Ramalingam, R., and Roth, A. (2023). Batch multivalid conformal prediction. In *International Conference on Learning Representations*.

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

Kivaranovic, D., Johnson, K. D., and Leeb, H. (2020). Adaptive, Distribution-Free Prediction Intervals for Deep Networks. In *International Conference on Artificial Intelligence and Statistics*. PMLR.

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).

Podkopaev, A. and Ramdas, A. (2021). Distribution-free uncertainty quantification for classification under label shift. In *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*. PMLR.

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. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc.

Sesia, M. and Romano, Y. (2021). Conformal prediction using conditional histograms. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc.

Tibshirani, R. J., Barber, R. F., Candes, E., and Ramdas, A. (2019). Conformal Prediction Under Covariate Shift. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc.

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

Wan, C., Xu, Z., Pinson, P., Dong, Z. Y., and Wong, K. P. (2014). Probabilistic forecasting of wind power generation using extreme learning machine. *IEEE Transactions on Power Systems*, 29(3):1033–1044.

Wang, H., Li, G., Wang, G., Peng, J., Jiang, H., and Liu, Y. (2017). Deep learning based ensemble approach for probabilistic wind power forecasting. *Applied Energy*, 188:56–70.

Wisniewski, W., Lindsay, D., and Lindsay, S. (2020). Application of conformal prediction interval estimations to market makers' net positions. Proceedings of Machine Learning Research. PMLR.

Zaffran, M., Féron, O., Goude, Y., Josse, J., and Dieuleveut, A. (2022). Adaptive conformal predictions for time series. In *Proceedings of the 39th International Conference on Machine Learning*. PMLR.

Time Series

Generalizing beyond exchangeability in theory

Two major general theoretical results beyond exchangeability:

- Chernozhukov et al. (2018)
  $\hookrightarrow$ If the learnt model is accurate and the data noise is strongly mixing, then CP is valid asymptotically ✓
- Barber et al. (2022)
  $\hookrightarrow$ Quantifies the coverage loss depending on the strength of exchangeability violation
  $$\mathbb{P}(Y_{n+1} \in \widehat{C}_\alpha(X_{n+1})) \geq 1 - \alpha - \underbrace{\text{average violation of exchangeability by each calibration point}}$$
  $\hookrightarrow$ proposed algorithm: reweighting (again)!
  e.g., in a temporal setting, give higher weights to more recent points.

Exchangeability does not hold in many practical applications

CP requires exchangeable data points to ensure validity

- ✗ Covariate shift, i.e. $\mathcal{L}_X$ changes but $\mathcal{L}_{Y|X}$ stays constant
  (see e.g., Tibshirani et al., 2019)
- ✗ Label shift, i.e. $\mathcal{L}_Y$ changes but $\mathcal{L}_{X|Y}$ stays constant
  (see e.g., Podkopaev and Ramdas, 2021)
- ✗ Arbitrary distribution shift
  (see e.g., Cauchois et al., 2020)
- ✗ Possibly many shifts, not only one
  (main focus of this presentation)

Recent developments

- Gibbs and Candès (2022) later on also proposes a method not requiring to choose γ
- Bhatnagar et al. (2023) enjoys **anytime** regret bound, by leveraging tools from the strongly adaptive regret minimization literature
- Feldman et al. (2023) extends ACI to general risk control
- Bastani et al. (2022) proposes an algorithm achieving stronger coverage guarantees (conditional on specified overlapping subsets, and threshold calibrated) without hold-out set
- Angelopoulos et al. (2023) combines CP ideas with control theory ones, to adaptively improve the predictive intervals depending on the errors structure

---

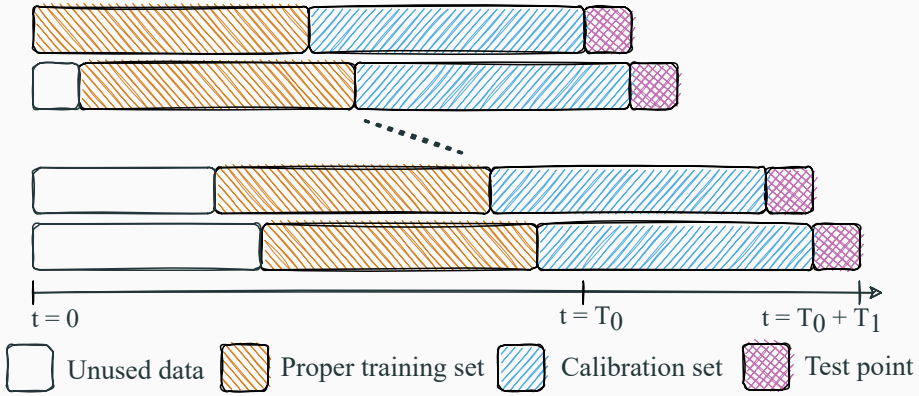
Non exhaustive references.

How to adapt to time series?

Usual ideas from the time series literature:

- Consider an online procedure (for each new data, re-train and re-calibrate)
  - ↪ update to recent observations (trend impact, period of the seasonality, dependence...)
- Use a sequential split
  - ↪ use only the past so as to correctly estimate the variance of the residuals (using the future leads to optimistic residuals and underestimation of their variance)

Online sequential split conformal prediction (OSSCP)



Wisniewski et al. (2020); Kath and Ziel (2021); Zaffran et al. (2022)

↪ tested on real time series

Time Series

- Literature on non-exchangeable CP
- Updating the training and calibration sets
- Theoretical analysis of ACI's length
- Numerical experiments

Missing Values

Numerical analysis of ACI's length: AR(1) case

Assume the residuals follow an AR(1) process: $\hat{\varepsilon}_{t+1} = \varphi\hat{\varepsilon}_t + \xi_{t+1}$ with $(\xi_t)_t$ i.i.d. random variables and other assumptions, we have:

$$\frac{1}{T}\sum_{t=1}^{T} L(\alpha_t) \xrightarrow[T\to+\infty]{a.s.} \mathbb{E}_{\pi_{\gamma,\varphi}}[L].$$

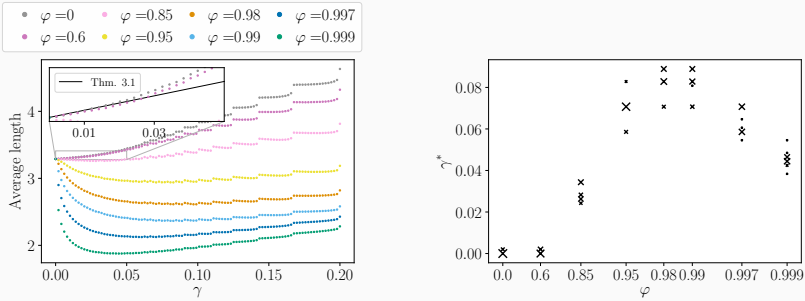


Figure 5: Left: evolution of the mean length depending on γ for various φ. Right: γ^* minimizing the average length for each φ.

Data generation and simulation settings

$$Y_t = 10 \sin(\pi X_{t,1} X_{t,2}) + 20 (X_{t,3} - 0.5)^2 + 10 X_{t,4} + 5 X_{t,5} + \varepsilon_t$$

where the $X_{t,\cdot} \sim \mathcal{U}([0,1])$ and ε_t is an ARMA(1,1) process:

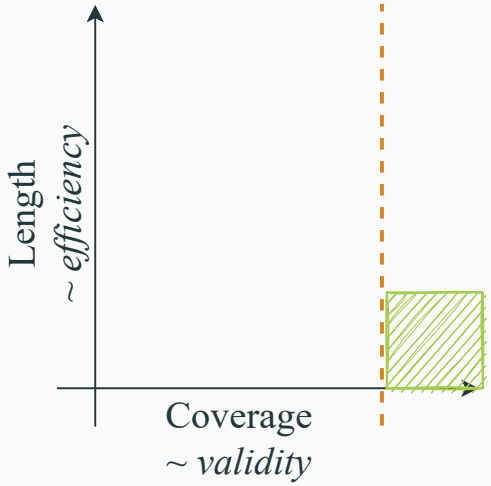
$$\varepsilon_{t+1} = \varphi \varepsilon_t + \xi_{t+1} + \theta \xi_t,$$

with ξ_t is a white noise of variance σ^2.

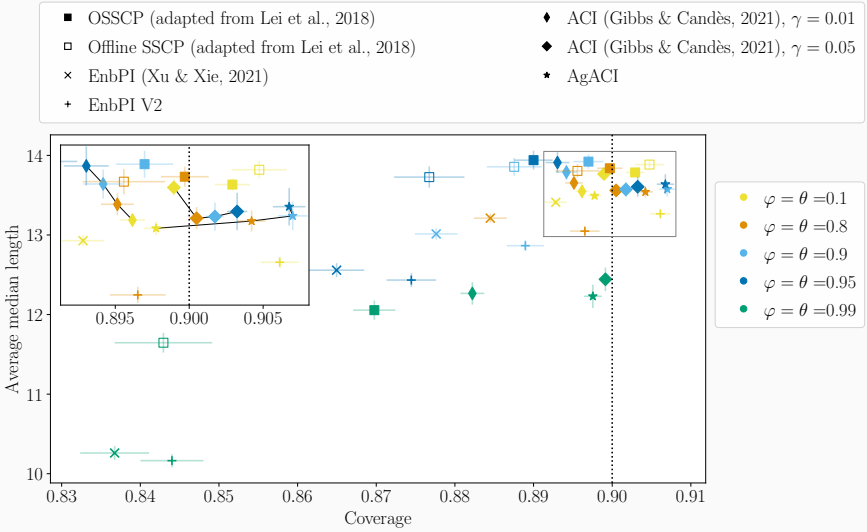
- $\varphi = \theta$ range in $[0.1, 0.8, 0.9, 0.95, 0.99]$.
- We fix σ to keep the variance $\mathrm{Var}(\varepsilon_t)$ constant to 10 (or 1).
- We use random forest as regressor.
- For each setting (pair variance and φ,θ):
  - 300 points, the last 100 kept for prediction and evaluation,
  - 500 repetitions,

  $\Rightarrow$ in total, $100 \times 500 = 50000$ predictions are evaluated.

Visualisation of the results



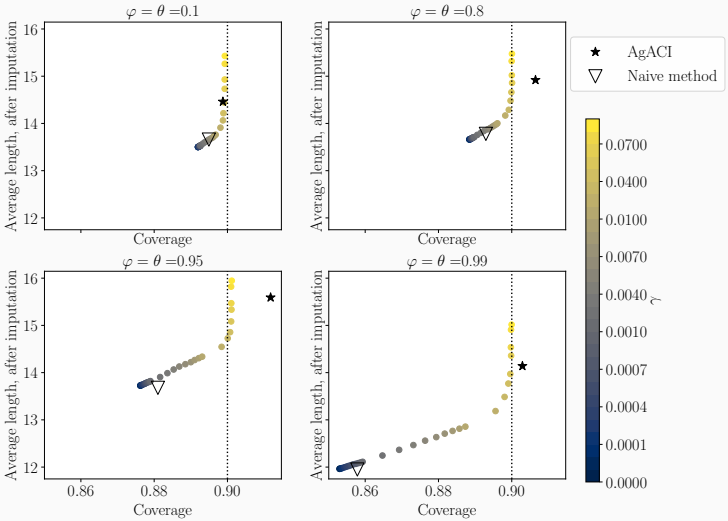
Results: impact of the temporal dependence, ARMA(1,1), variance 10



Summary

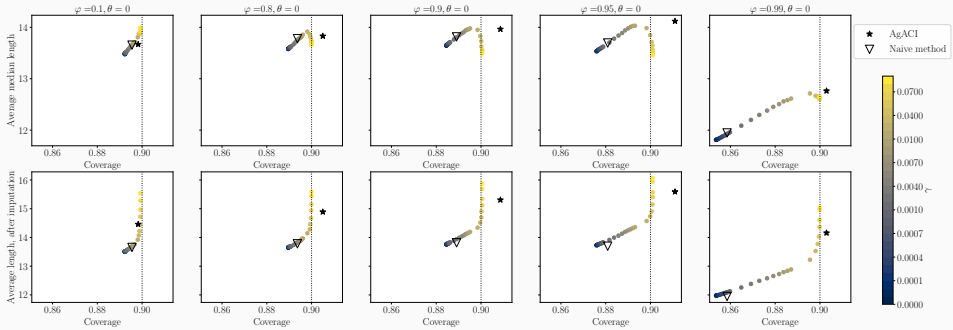
1. The temporal dependence impacts the *validity*.
2. Online is significantly better than offline.
3. **OSSCP.** Achieves *valid* coverage for φ and θ smaller than 0.9, but is not robust to the increasing dependence.
4. **EnbPI.** Its *validity* strongly depends on the data distribution. When the method is *valid*, it produces the smallest intervals. EnbPI V2 method should be preferred.
5. **ACI.** Achieves *valid* coverage for every simulation settings with a well chosen γ, or for dependence such that $\varphi < 0.95$. It is robust to the strength of the dependence.
6. **AgACI.** Achieves *valid* coverage for every simulation settings, with good *efficiency*.

Empirical evaluation of ACI sensitivity to γ and adaptive choice

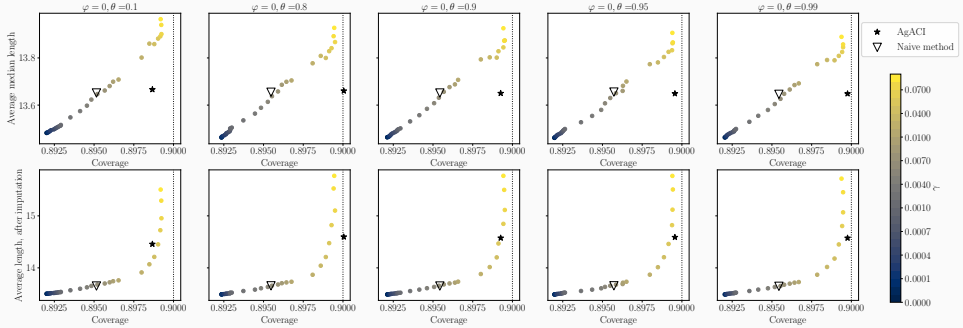


$\Rightarrow$ The more the dependence, the more sensitive to γ is ACI. Naive method ($\triangledown$): smallest among valid ones in the past $\Rightarrow$ accumulates error of the different ACI's versions. AgACI ($\bigstar$): encouraging preliminary results.

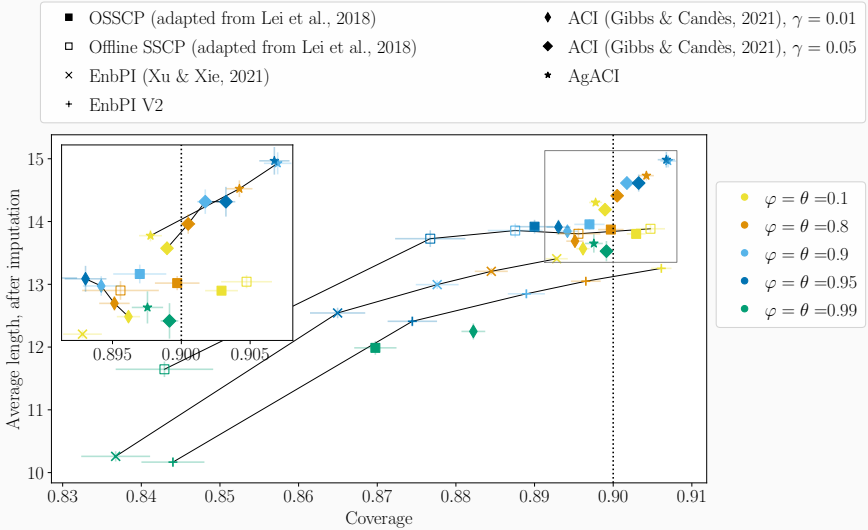
Empirical evaluation of ACI sensitivity to γ and adaptive choice, AR(1)



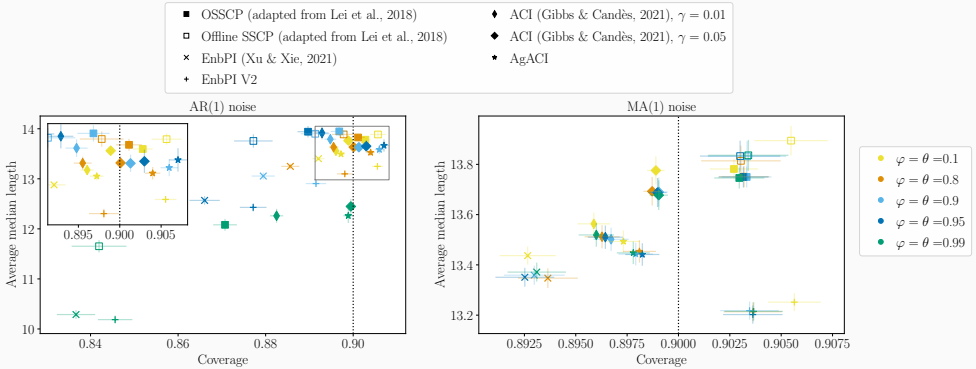
Empirical evaluation of ACI sensitivity to γ and adaptive choice, MA(1)



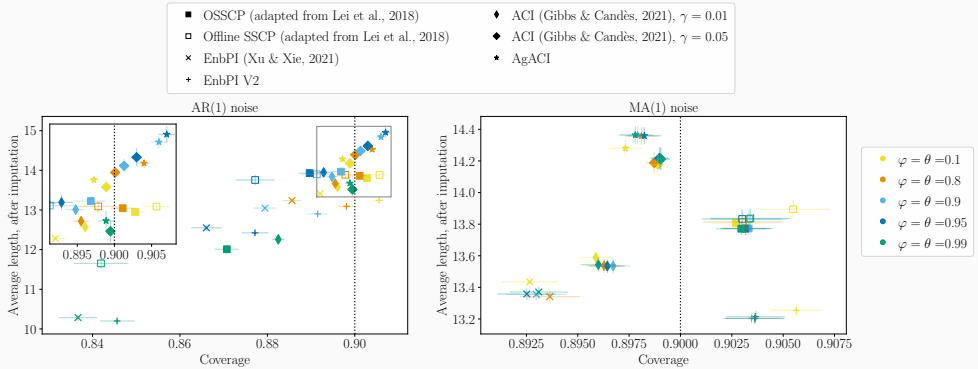
Results: impact of the temporal dependence, ARMA(1), variance 10, average length after imputation



Results: impact of the temporal dependence, AR(1) and MA(1), variance 10



Results: impact of the temporal dependence, AR(1) and MA(1), variance 10, average length after imputation



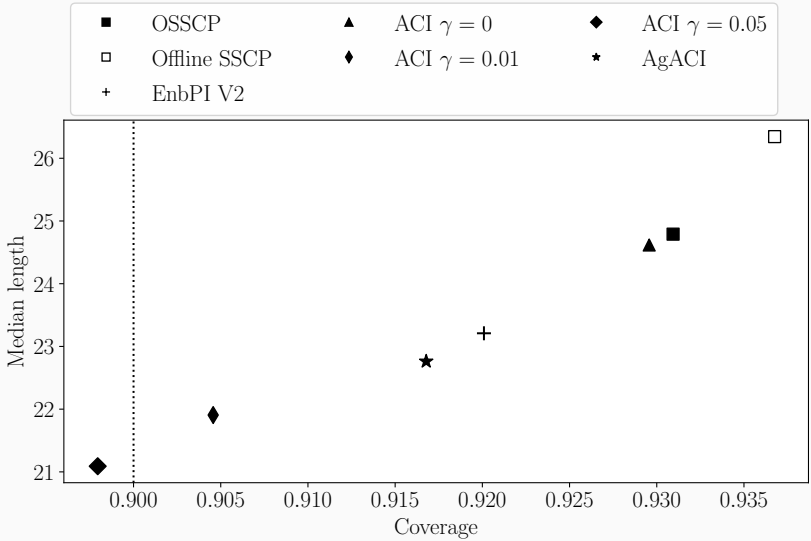
Forecasting electricity prices with confidence in 2019

- Forecast for the year 2019.
- Random forest regressor.
- One model per hour, we concatenate the predictions afterwards.

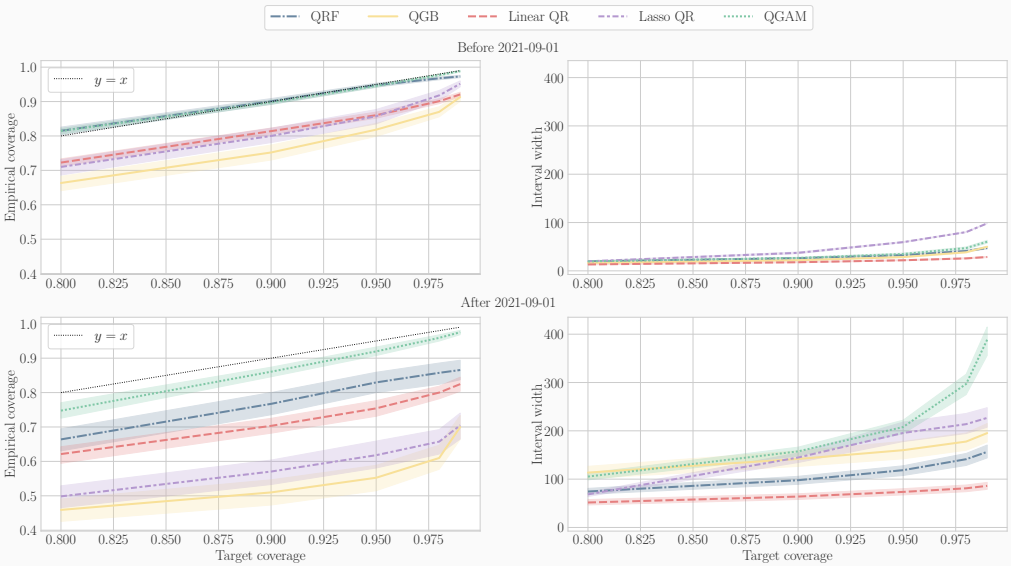
$\hookrightarrow$ 24 models

  - $y_t \in \mathbb{R}$
  - $x_t \in \mathbb{R}^d$, with $d = 24 + 24 + 1 + 7 = 56$
  - 3 years for training/calibration, i.e. $T_0 = 1096$ observations
  - 1 year to forecast, i.e. $T_1 = 365$ observations

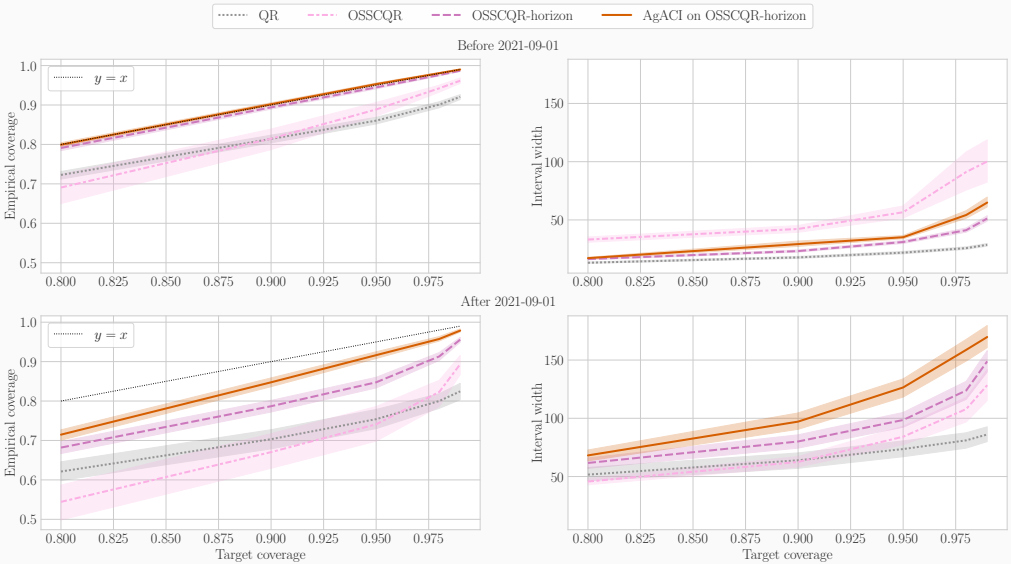
Performance on predicted French electricity Spot price for the year 2019

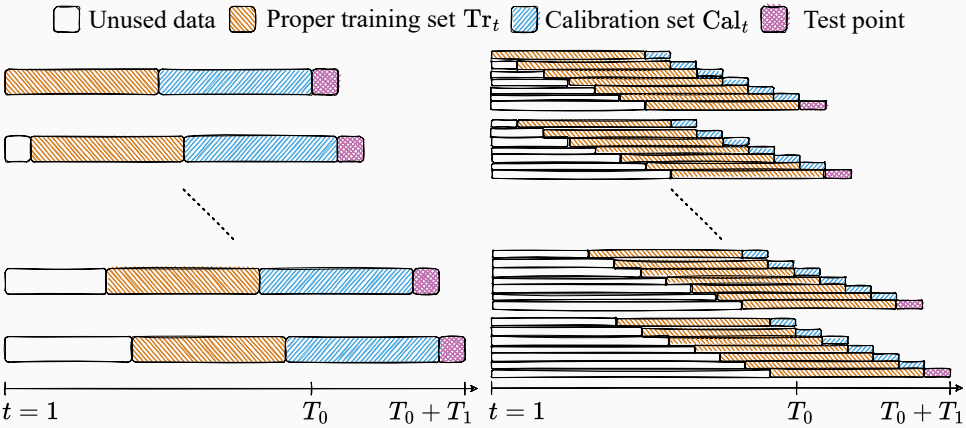


Forecasting electricity prices with confidence in 2020 and 2021, various models



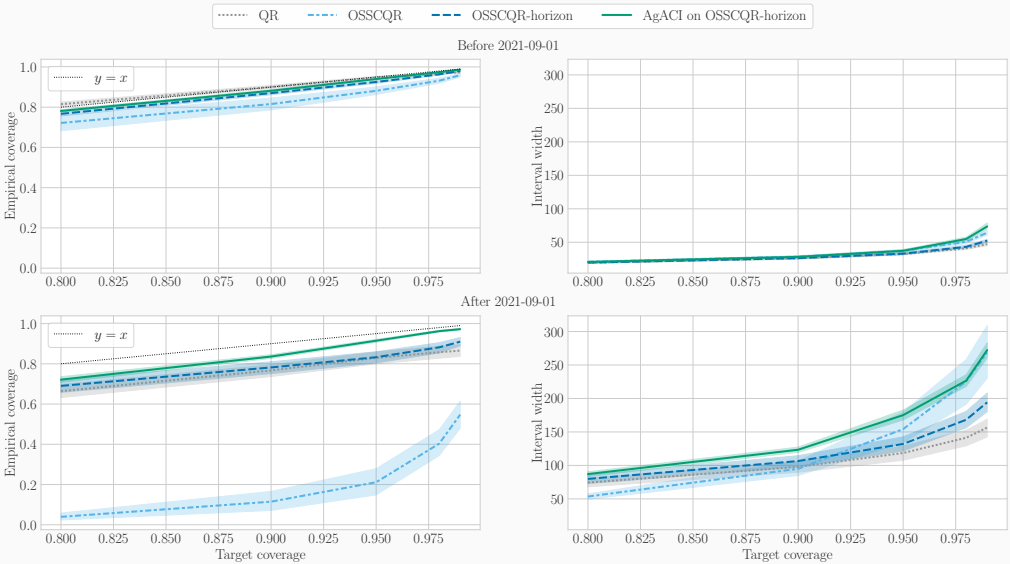
Forecasting electricity prices with confidence in 2020 and 2021, linear models



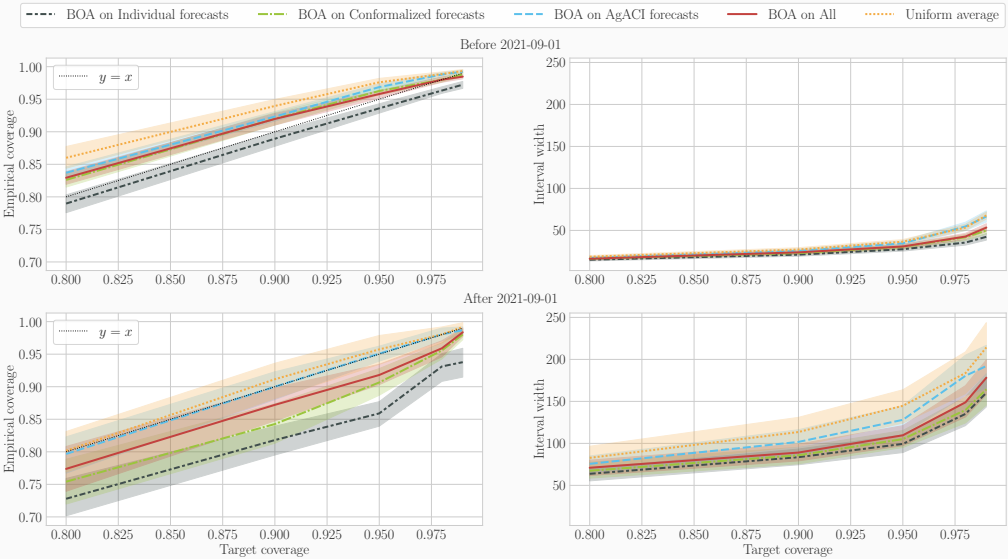


(a) OSSCP (b) OSSCP-horizon

Forecasting electricity prices with confidence in 2020 and 2021, QRF models



Forecasting electricity prices with confidence in 2020 and 2021, online aggregation models



Missing Values

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

Oracle int. under Gaussian lin. mod. **(Z., Dieuleveut, Josse, and Romano, 2023)**

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

- 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)**

Properties of isotonic predictive uncertainty

$$V(X_{\text{obs}(M)}, M) := \text{Var}\left(Y | X_{\text{obs}(M)}, M\right)$$

$$V(X_{\text{obs}(m)}, m) \overset{a.s.}{\leq} V(X_{\text{obs}(m')}, m') \qquad \text{for any } m \subset m', \tag{Var-1}$$

$$\mathbb{E}\left[V(X_{\text{obs}(M)}, M) | M = m\right] \leq \mathbb{E}\left[V(X_{\text{obs}(M)}, M) | M = m'\right] \qquad \text{for any } m \subset m'. \tag{Var-2}$$

$$\text{IQ}_{\beta,\gamma}(X_{\text{obs}(m)}, m) \overset{a.s.}{\leq} \text{IQ}_{\beta,\gamma}(X_{\text{obs}(m')}, m') \qquad \text{for any } m \subset m', \tag{IQ-1}$$

$$\mathbb{E}\left[\text{IQ}_{\beta,\gamma}(X_{\text{obs}(M)}, M) | M = m\right] \leq \mathbb{E}\left[\text{IQ}_{\beta,\gamma}(X_{\text{obs}(M)}, M) | M = m'\right] \qquad \text{for any } m \subset m'. \tag{IQ-2}$$

$$\Lambda(\mathcal{C}_\alpha(X_{\text{obs}(m)}, m)) \overset{a.s.}{\leq} \Lambda(\mathcal{C}_\alpha(X_{\text{obs}(m')}, m')) \qquad \text{for any } m \subset m', \tag{Len-1}$$

$$\mathbb{E}\left[\Lambda(\mathcal{C}_\alpha(X_{\text{obs}(M)}, M)) | M = m\right] \leq \mathbb{E}\left[\Lambda(\mathcal{C}_\alpha(X_{\text{obs}(M)}, M)) | M = m'\right] \qquad \text{for any } m \subset m'. \tag{Len-2}$$

Results

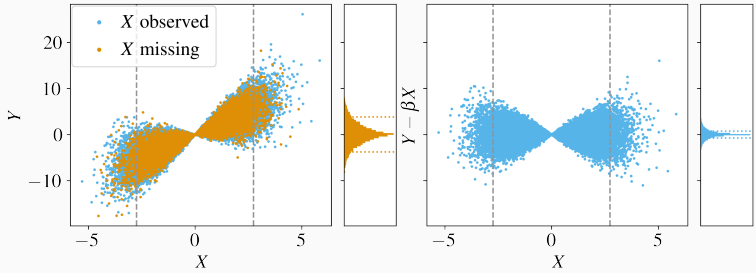
Property \ Setup	GLM homoske.	GLM heteroske.	$M \perp\!\!\!\perp X$ and $Y \perp\!\!\!\perp M \mid X$
Variance	Var-1	~~Var-1~~ Var-2	Var-2
Inter-quantile	IQ-1	IQ-2	
Length of Oracle PI	Len-1	Len-2	Len-2

Univariate heteroskedastic Gaussian linear model

Unidimensional heteroskedasticity

Consider the following one-dimensional model:

- $X \sim \mathcal{N}(0, \sigma^2)$, $\sigma \in \mathbb{R}_+$;
- $\xi \sim \mathcal{N}(0, \tau^2)$, $\tau \in \mathbb{R}_+$, such that $\xi \perp\!\!\!\perp X$;
- $Y = \beta X + X\xi$, with $\beta \in \mathbb{R}$;
- $M \sim \mathcal{B}(\rho)$, with $\rho \in [0, 1]$, and $M \perp\!\!\!\perp (X, Y)$.



Input: *i)* Training set $\left\{\left(X^{(k)}, M^{(k)}, Y^{(k)}\right)\right\}_{k=1}^{n}$. *ii)* imputation algorithm $\mathcal{I}$. *iii)* learning algorithm $\mathcal{A}$ taking its values in $\mathcal{F} := \mathcal{Y}^{\mathcal{X}\times\mathcal{M}}$. *iv)* calibration proportion $\rho \in]0, 1]$. *v)* $\left\{\mathrm{Tr}, \mathrm{Cal}, \Phi, \hat{A}\right\}$ the output of the splitting algorithm 1 ran on $\left\{\left\{\left(X^{(k)}, M^{(k)}, Y^{(k)}\right)\right\}_{k=1}^{n}, \mathcal{I}, \mathcal{A}, \rho\right\}$. *vi)* conformity score function $s(\cdot, \cdot; f)$ for $f \in \mathcal{F}$. *vii)* significance level α. *viii)* test point $\left(X^{(n+1)}, M^{(n+1)}\right)$. *ix)* subsampled set of calibration indices $\widetilde{\mathrm{Cal}} \subseteq \mathrm{Cal}$ **for** $k \in \widetilde{\mathrm{Cal}}$: $\widetilde{M}^{(k)} = \max(M^{(k)}, M^{(n+1)})$

$$\widehat{C}_{\alpha}^{\text{MDA-Nested}^\star}\left(X^{(n+1)}, M^{(n+1)}\right) := \left\{y \in \mathcal{Y} : (1-\alpha)(1+\#\widetilde{\mathrm{Cal}}) > \sum_{k\in\mathrm{Cal}} \mathbb{1}\left\{s\left(\left(X^{(k)}, \widetilde{M}^{(k)}\right), Y^{(k)}; \hat{A}\left(\Phi\left(\cdot,\cdot\right),\cdot\right)\right) < s\left(\left(X^{(n+1)}, \widetilde{M}^{(k)}\right), y; \hat{A}\left(\Phi\left(\cdot,\cdot\right),\cdot\right)\right)\right\}\right\}$$

Algorithm 1 Split and train

Input: Imputation algorithm $\mathcal{I}$, learning algorithm $\mathcal{A}$ taking its values in $\mathcal{F} := \mathcal{Y}^{\mathcal{X}\times\mathcal{M}}$, training set $\left\{\left(X^{(k)}, M^{(k)}, Y^{(k)}\right)\right\}_{k=1}^{n}$, calibration proportion $\rho \in]0,1]$

Output: Splitted sets of indices Tr and Cal, imputation function Φ, fitted predictor $\hat{A}$

1: Randomly split $\{1,\ldots,n\}$ into 2 disjoint sets Tr & Cal of sizes $\#\mathrm{Tr} = (1-\rho)n$ and $\#\mathrm{Cal} = \rho n$

2: Fit the imputation function: $\Phi(\cdot,\cdot) \leftarrow \mathcal{I}\left(\left\{\left(X^{(k)}, M^{(k)}\right), k \in \mathrm{Tr}\right\}\right)$

3: Fit the learning algorithm $\mathcal{A}$: $\hat{A}(\cdot,\cdot) \leftarrow \mathcal{A}\left(\left\{\left(\Phi\left(X^{(k)}, M^{(k)}\right), M^{(k)}\right), k \in \mathrm{Tr}\right\}\right)$

CP-MDA-Nested* is Marginally Valid (MV)

CP-MDA-Nested* marginal validity (Z., Josse, Romano, and Dieuleveut, 2024)

Under the assumptions that:

- $\left(X^{(k)}, M^{(k)}, Y^{(k)}\right)_{k=1}^{n+1}$ are exchangeable,
- the subsampling scheme keeps all of the calibration points,

then, for almost all imputation function, CP-MDA-Nested* reaches (MV) at the level $1-2\alpha$, that is:

$$\mathbb{P}_{\mathcal{D}^{\text{exch}(n+1)}}\left\{Y^{(n+1)} \in \widehat{C}_\alpha\left(X^{(n+1)}, M^{(n+1)}\right)\right\} \geq 1-2\alpha.$$

- ✓ 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

MDA-Exact achieves Mask-Conditional-Validity (MCV)

`CP-MDA-Exact` achieves exact MCV (Z., Dieuleveut, Josse, and Romano, 2023)

If:

- $\left(X^{(k)}, M^{(k)}, Y^{(k)}\right)_{k=1}^{n+1}$ are i.i.d.,
- $M \perp\!\!\!\perp X$,
- $Y \perp\!\!\!\perp M \,|X$,

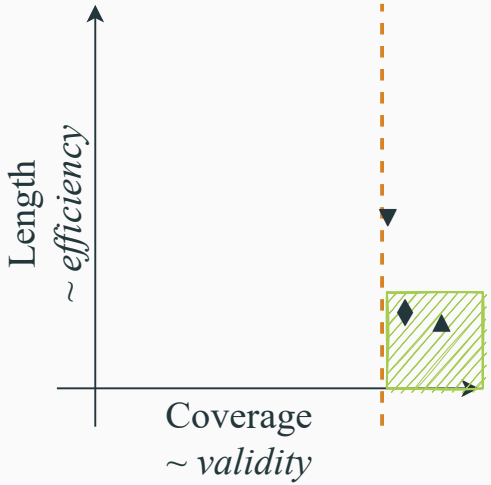
then, for almost all imputation function, `CP-MDA-Exact` is such that for any $m \in \{0,1\}^d$ such that $\mathbb{P}_{\mathcal{D}}(M = m) > 0$:

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

and if additionally the scores are almost surely distinct:

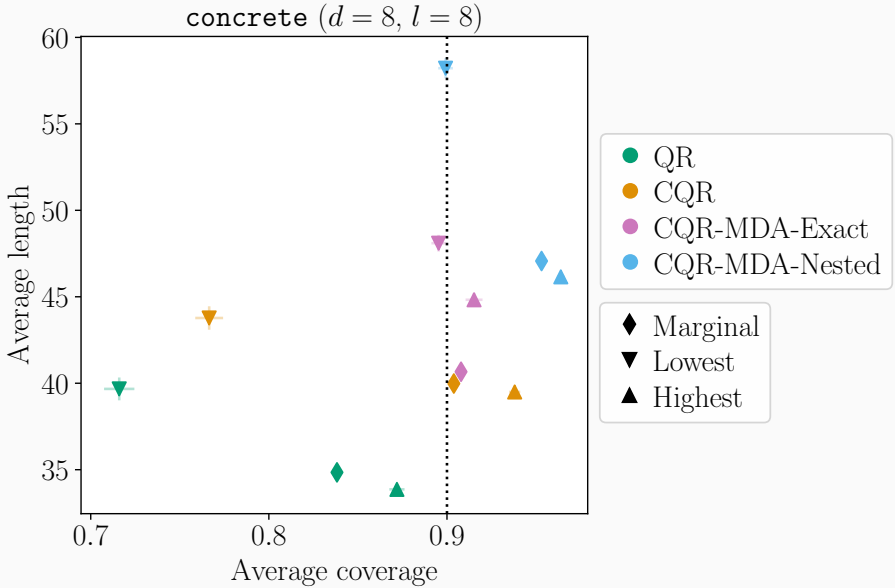
$$\mathbb{P}_{\mathcal{D}^{\otimes(n+1)}}\left\{Y^{(n+1)} \in \widehat{C}_\alpha\left(X^{(n+1)}, M^{(n+1)}\right) | M^{(n+1)} = m\right\} \leq 1 - \alpha + \frac{1}{\#\mathrm{Cal}^m + 1}.$$

Before more experiments, visualisation

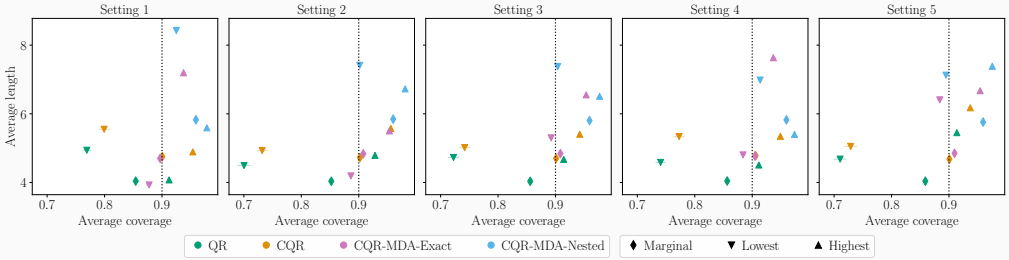


- ♦ : marginal coverage, i.e.
  $\mathbb{P}(Y \in \hat{C}_\alpha(X, M))$
- ▼ : lowest coverage, i.e.
  $\min_{m \in \mathcal{M}} \mathbb{P}(Y \in \hat{C}_\alpha(X, m) | M = m)$
- ▲ : highest coverage, i.e.
  $\max_{m \in \mathcal{M}} \mathbb{P}(Y \in \hat{C}_\alpha(X, m) | M = m)$

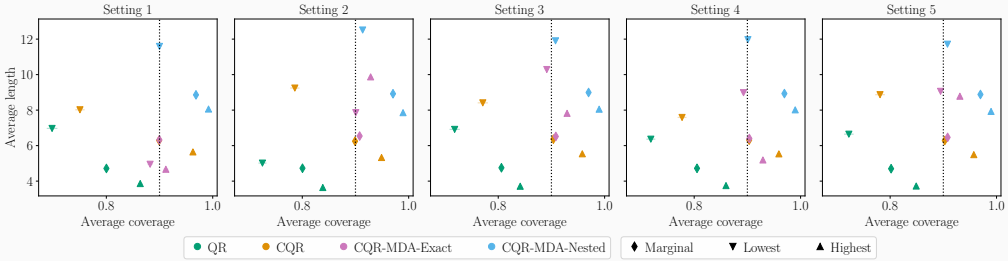
Semi-synthetic experiments



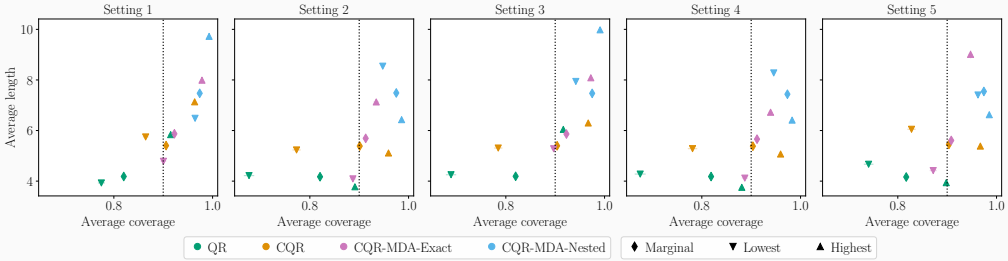
MAR, correlation coefficient of 0.8



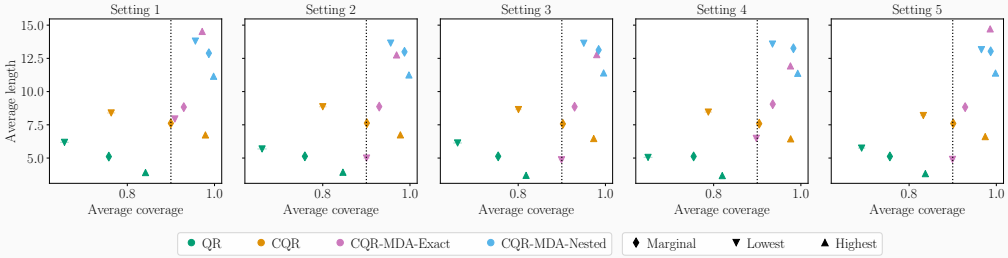
MAR, independent features



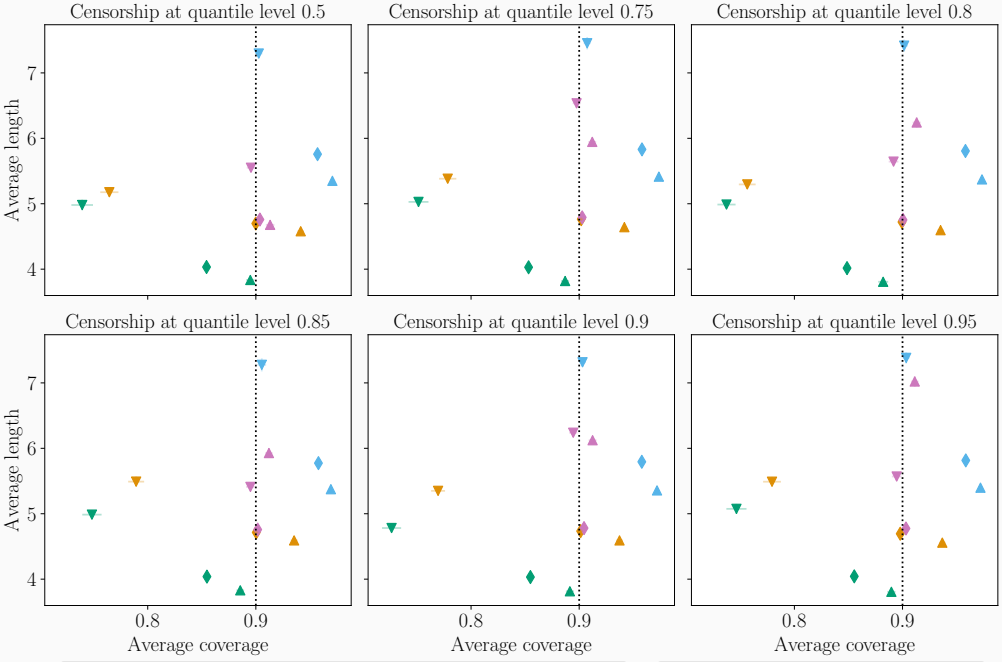
MNAR self-masked, correlation coefficient of 0.8



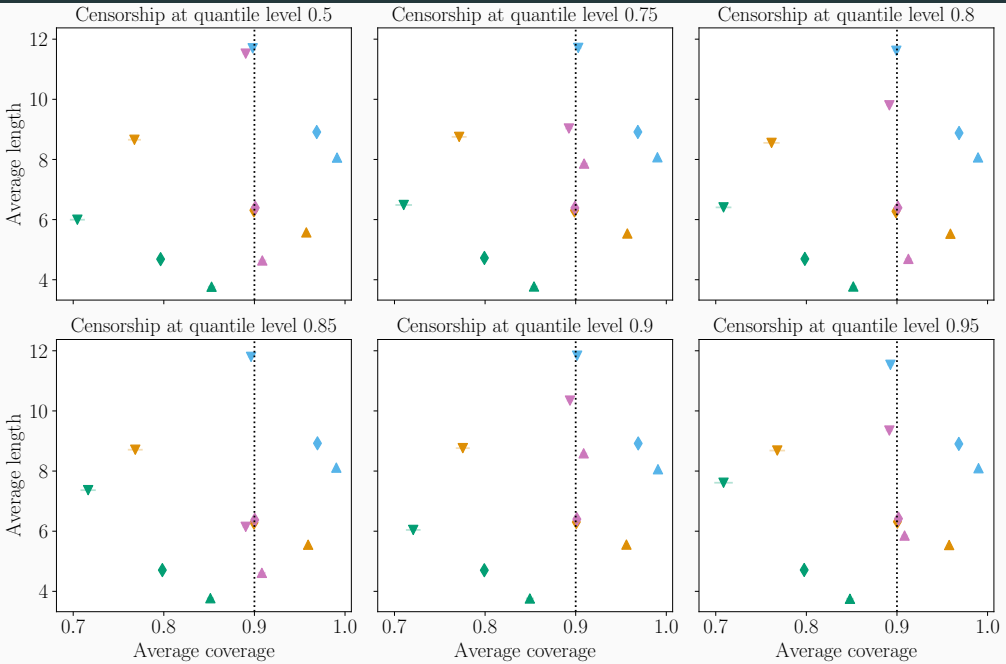
MNAR self-masked, independent features



MNAR quantile censorship, correlation coefficient of 0.8

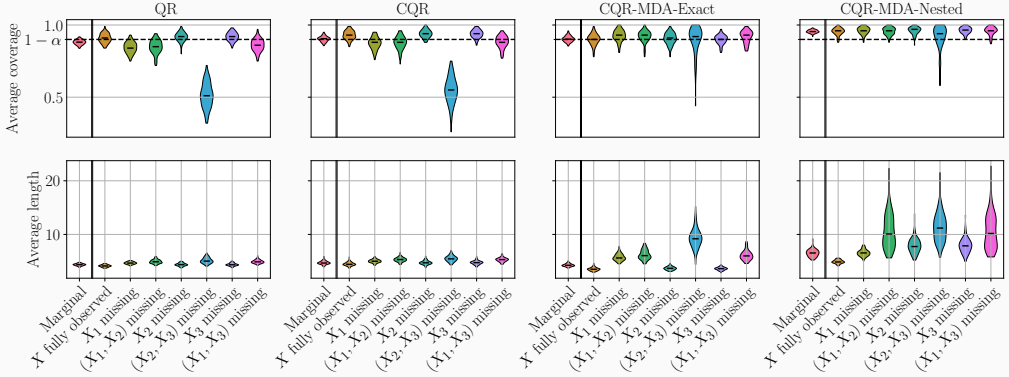


MNAR quantile censorship, independent features



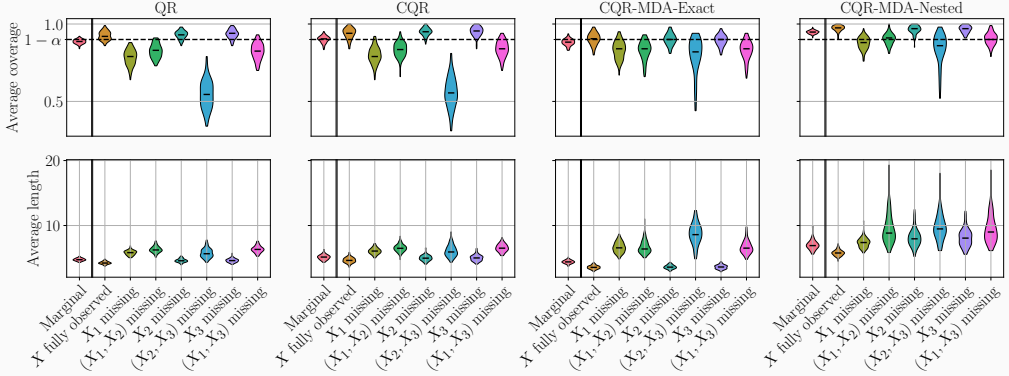
$Y \not\perp\!\!\!\perp M \,|X$, correlation coefficient of 0.8 ($d = 3$)

- $\varepsilon \sim \mathcal{N}(0,1) \perp\!\!\!\perp (X, M)$,
- $X \sim \mathcal{N}(\mu, \Sigma)$, $\mu = (1,1,1)^T$, $\Sigma = \varphi(1,1,1)^T(1,1,1) + (1-\varphi)I_d$, $\varphi = 0.8$,
- $M_i \sim \mathcal{B}(0.2)$ for any $i \in [\![1,3]\!]$, independently from X and ε,
- $Y = X_1 \mathbb{1}\{M_1 = 0\} + 2X_1 \mathbb{1}\{M_1 = 1\} + 3X_2 \mathbb{1}\{M_2 = 1, M_3 = 1\} + \varepsilon$.



$Y \not\perp\!\!\!\perp M \,|X$, independent features ($d = 3$)

- $\varepsilon \sim \mathcal{N}(0,1) \perp\!\!\!\perp (X, M)$,
- $X \sim \mathcal{N}(\mu, \Sigma)$, $\mu = (1,1,1)^T$, $\Sigma = \varphi(1,1,1)^T(1,1,1) + (1-\varphi)I_d$, $\varphi = 0$,
- $M_i \sim \mathcal{B}(0.2)$ for any $i \in [\![1,3]\!]$, independently from X and ε,
- $Y = X_1 \mathbb{1}\{M_1 = 0\} + 2X_1 \mathbb{1}\{M_1 = 1\} + 3X_2 \mathbb{1}\{M_2 = 1, M_3 = 1\} + \varepsilon$.



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 blood platelets upon arrival at hospital, given 7 pre-hospital features.

These covariates are not always observed.

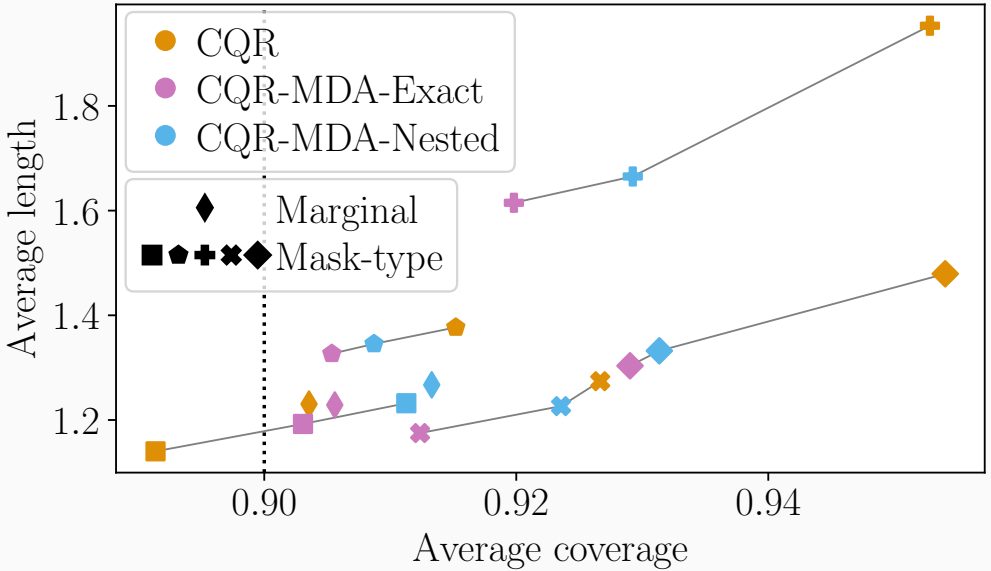
Data set description i

- `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 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).

Real data experiment: TraumaBase®, critical care medicine



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)*.

Pinball-consistency of an universal learner **(Z., Dieuleveut, Josse, and Romano, 2023)**

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 **?**.

Asymptotic conditional coverage of a universal quantile learner

Corollary (Z., Dieuleveut, Josse, and Romano, 2023)

For any missing mechanism, for almost all $\mathcal{C}^\infty$ imputation function Φ, if $F_{Y|(X_{\text{obs}(\text{M})},M)}$ 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.