

Conformal Prediction with Missing Values

Margaux Zaffran

StatMathAppli 2023





Aymeric Dieuleveut

Ecole
Polytechnique
Paris - France



Julie Josse

INRIA
IDESP
Montpellier - France



Yaniv Romano

Technion - Israel Institute
of Technology
Haifa - Israel

(Way too short) Intro to (Split) Conformal Prediction

Standard Split Conformal Prediction for Mean-Regression

Improving Adaptiveness: Conformalized Quantile Regression

Conformal Prediction with Missing Values

- $(X, Y) \in \mathbb{R}^d \times \mathbb{R}$ random variables
- n training samples $(X^{(k)}, Y^{(k)})_{k=1}^n$
- **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}_α such that:

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

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

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

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

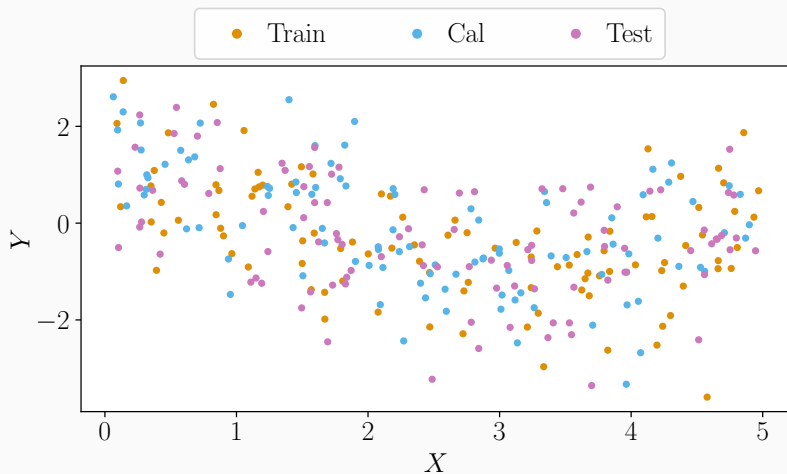
(Way too short) Intro to (Split) Conformal Prediction

Standard Split Conformal Prediction for Mean-Regression

Improving Adaptiveness: Conformalized Quantile Regression

Conformal Prediction with Missing Values

Split Conformal Prediction (SCP)^{1,2,3}: toy example

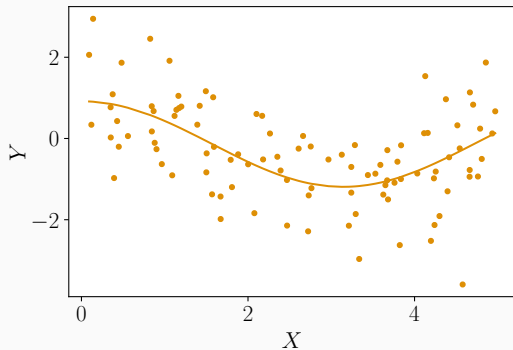


¹Vovk et al. (2005), *Algorithmic Learning in a Random World*

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

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

Split Conformal Prediction (SCP)^{1,2,3}: training step



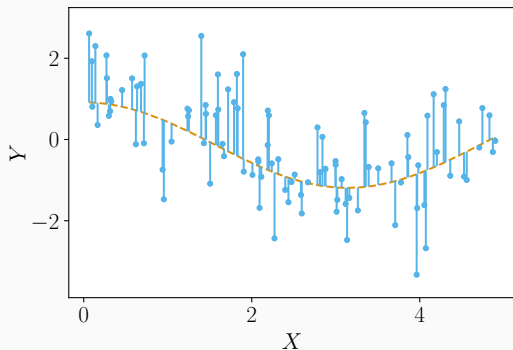
► Learn (or get) $\hat{\mu}$

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

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

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

Split Conformal Prediction (SCP)^{1,2,3}: calibration step



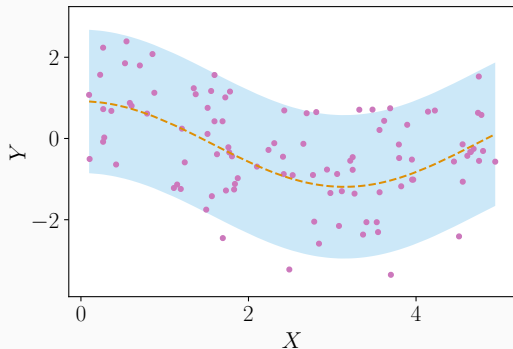
- ▶ Predict with $\hat{\mu}$
- ▶ Get the **|residuals|**, a.k.a. scores $\{S^{(k)}\}_{k \in \text{Cal}}$
- ▶ Compute the $(1 - \alpha)$ empirical quantile of $\mathcal{S} = \{|\text{residuals}|\}_{\text{Cal}} \cup \{+\infty\}$, noted $q_{1-\alpha}(\mathcal{S})$

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

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

³Lei 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 $\hat{C}_\alpha(x)$: $[\hat{\mu}(x) \pm q_{1-\alpha}(\mathcal{S})]$

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

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

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

Definition (Exchangeability)

$(X^{(k)}, Y^{(k)})_{k=1}^n$ are **exchangeable** if for any permutation σ of $\llbracket 1, n \rrbracket$ we have:

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

where \mathcal{L} designates the joint distribution.

Examples of exchangeable sequences

- i.i.d. samples

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

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

Theorem

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

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

Additionally, if the scores $\{S^{(k)}\}_{k \in \text{Cal}} \cup \{S_{n+1}\}$ are a.s. distinct:

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

- ✓ Distribution-free, only requires exchangeability
- ✓ Any regression algorithm (neural nets, random forest...)
- ✓ Finite sample
- ✗ Marginal coverage: $\mathbb{P} \left\{ Y^{(n+1)} \in \widehat{C}_\alpha \left(X^{(n+1)} \right) \mid X^{(n+1)} = x \right\} \geq 1 - \alpha$

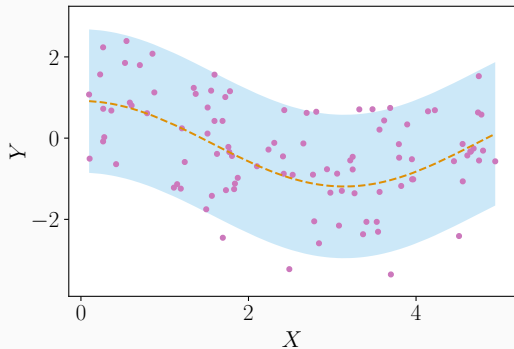
(Way too short) Intro to (Split) Conformal Prediction

Standard Split Conformal Prediction for Mean-Regression

Improving Adaptiveness: Conformalized Quantile Regression

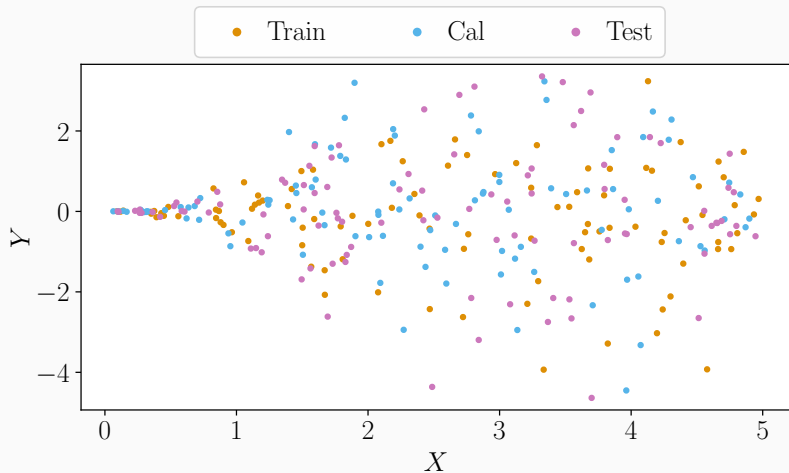
Conformal Prediction with Missing Values

Standard mean-regression SCP is not adaptive



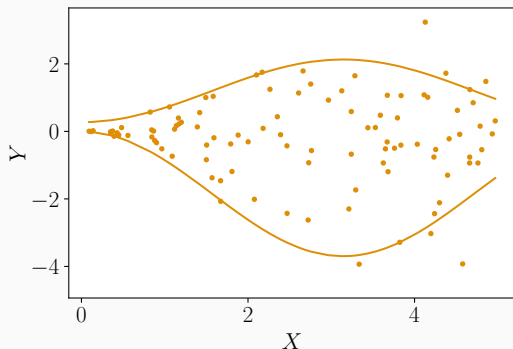
- ▶ Predict with $\hat{\mu}$
- ▶ Build $\hat{C}_\alpha(x)$: $[\hat{\mu}(x) \pm q_{1-\alpha}(\mathcal{S})]$

Conformalized Quantile Regression (CQR)⁴



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

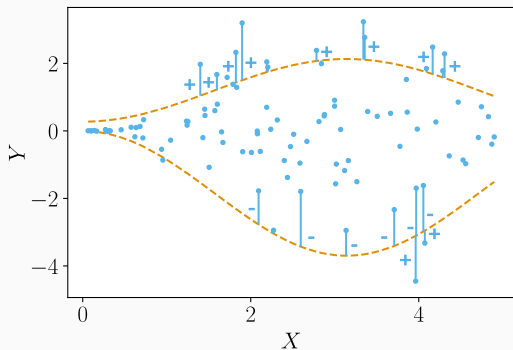
Conformalized Quantile Regression (CQR)⁴: training step



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

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

Conformalized Quantile Regression (CQR)⁴: calibration step

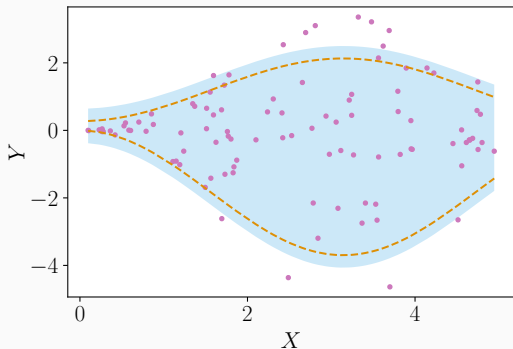


- ▶ Predict with $\widehat{QR}_{\text{lower}}$ and $\widehat{QR}_{\text{upper}}$
- ▶ Get the scores $\mathcal{S} = \{S^{(k)}\}_{\text{Cal}} \cup \{+\infty\}$
- ▶ Compute the $(1 - \alpha)$ empirical quantile of \mathcal{S} , noted $q_{1-\alpha}(\mathcal{S})$

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

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

Conformalized Quantile Regression (CQR)⁴: prediction step



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

► Build

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

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

CQR: theoretical guarantees

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

Theorem

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

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

Additionally, if the scores $\{S^{(k)}\}_{k \in \text{Cal}} \cup \{S_{n+1}\}$ are a.s. distinct:

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

- ✓ Distribution-free, only requires exchangeability
- ✓ Any quantile regression algorithm (neural nets, random forest...)
- ✓ Finite sample
- ✗ Marginal coverage: $\mathbb{P} \left\{ Y^{(n+1)} \in \widehat{C}_\alpha \left(X^{(n+1)} \right) \mid X^{(n+1)} = x \right\} \geq 1 - \alpha$ conditional

(Way too short) Intro to (Split) Conformal Prediction

Conformal Prediction with Missing Values

Missing values are ubiquitous and challenging

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

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

↔ 2^d potential masks.

↔ M can depend on X or Y (depending on the missing mechanism).

⇒ 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 .

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

$\xrightarrow{\phi}$

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

2. Train your algorithm (Random Forest, Neural Nets, etc.) on the **imputed**

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

↪ we consider an **impute-then-regress** pipeline in this work.

Predictive uncertainty quantification with missing values

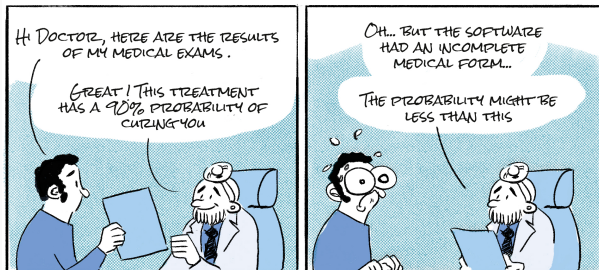
Goal: predict $Y^{(n+1)}$ with **confidence** $1 - \alpha$, i.e. build the smallest C_α such that:

1. Marginal Validity (MV)

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

2. Mask-Conditional-Validity (MCV)

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



CP is marginally valid (MV) after imputation

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

Lemma

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

Then, for *any missing mechanism*, for almost *all imputation function*⁵ ϕ :

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

\Rightarrow CQR, and Conformal Prediction, applied on an imputed data set still enjoys marginal guarantees⁶:

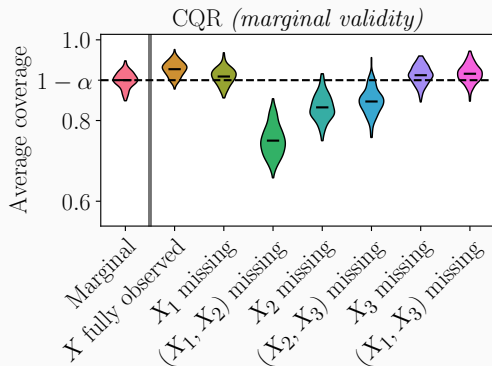
$$\mathbb{P} \left\{ Y^{(n+1)} \in \widehat{C}_\alpha \left(X^{(n+1)}, M^{(n+1)} \right) \right\} \geq 1 - \alpha.$$

⁵Even if the imputation is not accurate, the guarantee will hold.

⁶The upper bound also holds under continuously distributed scores.

CQR is marginally valid on imputed data sets

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

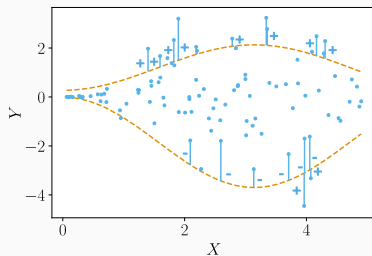


- The predictive uncertainty strongly depends on the mask

	Imputation+CQR	
(MV)	✓	
(MCV)	✗	

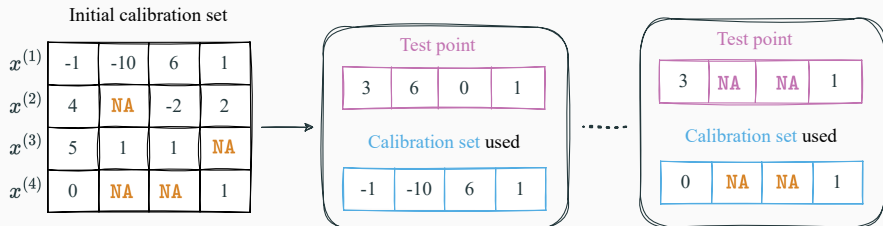
Conformalization step is independent of the important variable: the mask!

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



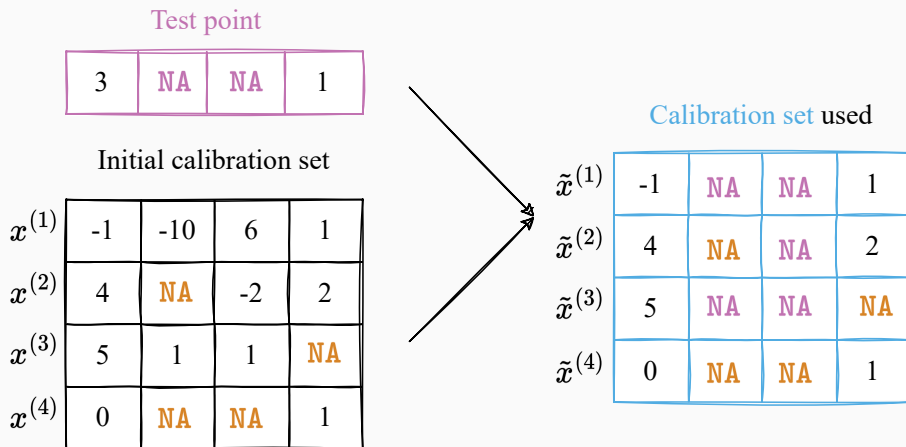
Warning: 2^d possible masks

⇒ Splitting the calibration set by mask is infeasible (lack of data)!



Missing Data Augmentation (MDA)

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

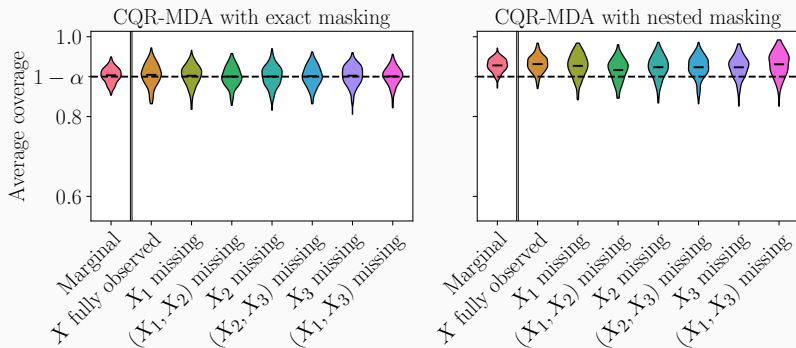


Algorithms: MDA with **Exact** masking or with **Nested** masking.

MDA achieves Mask-Conditional-Validity (MCV)

Theorem (Informal)

If $M \perp\!\!\!\perp (X, Y)$, for almost all imputation function, CP-MDA reaches (MCV).



	Imputation+CQR	CQR-MDA
(MV)	✓	✓
(MCV)	✗	✓

Questions? :)

Thanks for listening and feel free to reach out!

Paper →

Code →

Summary →



- Barber, R. F., Candès, E. J., Ramdas, A., and Tibshirani, R. J. (2021). The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*, 10(2).
- Chernozhukov, V., Wüthrich, K., and Zhu, Y. (2021). Distributional conformal prediction. *Proceedings of the National Academy of Sciences*, 118(48).
- Gibbs, I., Cherian, J. J., and Candès, E. J. (2023). Conformal prediction with conditional guarantees. arXiv: 2305.12616.
- Guan, L. (2022). Localized conformal prediction: a generalized inference framework for conformal prediction. *Biometrika*, 110(1).
- 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*.
- 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., 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).
- Papadopoulos, H., Proedrou, K., Vovk, V., and Gammerman, A. (2002). Inductive Confidence Machines for Regression. In *Machine Learning: ECML*. Springer.

- 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.
- Vovk, V. (2012). Conditional Validity of Inductive Conformal Predictors. In *Asian Conference on Machine Learning*. PMLR.
- Vovk, V., Gammerman, A., and Shafer, G. (2005). *Algorithmic Learning in a Random World*. Springer US.

Informative conditional coverage as such is impossible

- Impossibility results

↪ Lei and Wasserman (2014); Vovk (2012); Barber et al. (2021)

Without distribution assumption, in finite sample, a perfectly **conditionally valid** \widehat{C}_α is such that $\mathbb{P} \left\{ \text{mes} \left(\widehat{C}_\alpha(x) \right) = \infty \right\} = 1$ for any non-atomic x .

- Approximate conditional coverage

↪ Romano et al. (2020); Guan (2022); Jung et al. (2023); Gibbs et al. (2023)

Target $\mathbb{P}(Y_{n+1} \in \widehat{C}_\alpha | X_{n+1} \in \mathcal{R}(x)) \geq 1 - \alpha$

- Asymptotic (with the sample size) conditional coverage

↪ Romano et al. (2019); Kivaranovic et al. (2020); Chernozhukov et al. (2021); Sesia and Romano (2021); Izbicki et al. (2022)

CP-MDA with Exact masking

Test point

3	NA	NA	1
---	----	----	---

Initial calibration set

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



Calibration set used

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

#Cal^{M(test)} observations

CQR-MDA with exact masking in words

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

5.1 For each $j \in \llbracket 1, d \rrbracket$ s.t. $M_j^{(n+1)} = 1$, set $\tilde{M}_j^{(k)} = 1$
for k in **Cal** s.t. $M^{(k)} \subset M^{(n+1)}$

5.2 Impute the new **calibration set**

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

5.4 Impute the **test point**

5.5 Predict with the **quantile regressors** and the **correction** previously obtained,
 $q_{1-\alpha}(\mathcal{S})$

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

MDA achieves (MCV) in an informative way

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

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

