On the hardness of group-conditional distribution-free predictive inference

An application to prediction with missing covariates

Margaux Zaffran

February 6, 2025

Statistics Seminar, Laboratoire de Mathématiques d'Orsay

Julie Josse
PreMeDICaL
Inria

Yaniv Romano
Technion – Israel Institute of Technology

Aymeric Dieuleveut
CMAP
École Polytechnique

- *Predictive Uncertainty Quantification with Missing Covariates*, 2024
- *Conformal Prediction with Missing Values*, ICML 2023

Distribution-free predictive uncertainty quantification

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.

▶ *Validity* should be ensured

- in finite samples
- for all data distribution and underlying learnt model

Distribution-free marginal validity is achievable

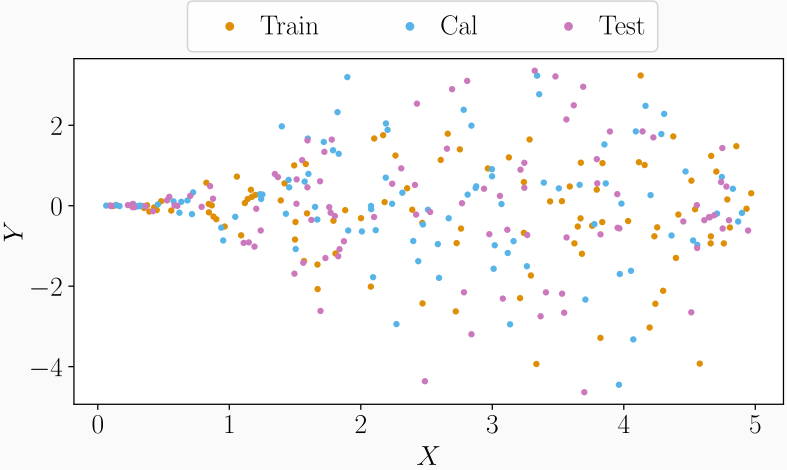
Conformal Prediction (Vovk et al., 2005; Papadopoulos et al., 2002; Lei et al., 2018) builds an estimated predictive set $\widehat{C}_\alpha$ based on n data points.

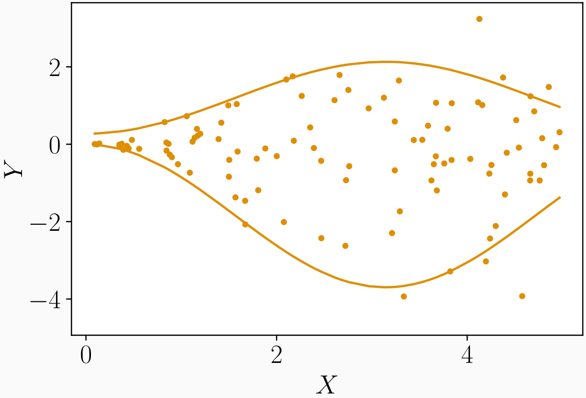
Conformal prediction achieves marginal validity (Vovk et al., 2005)

$\widehat{C}_\alpha$ outputted by conformal prediction is such that for any distribution $\mathcal{D}$ on $(\mathcal{X}, \mathcal{Y})$, it holds:

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

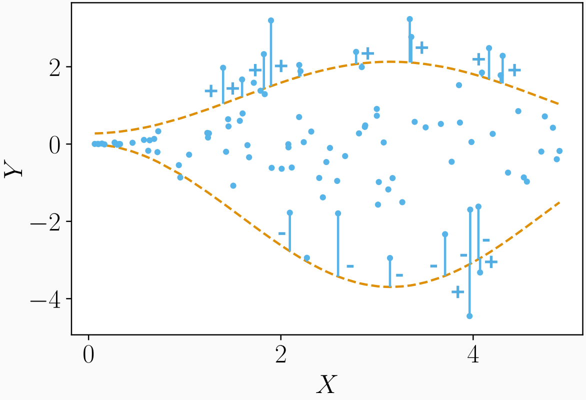
Conformalized Quantile Regression (CQR, Romano et al., 2019)





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

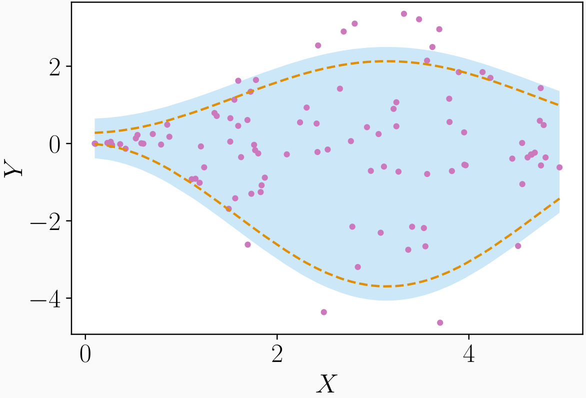
Conformalized Quantile Regression (CQR, Romano et al., 2019) calibration step



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

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

Conformalized Quantile Regression (CQR, Romano et al., 2019) prediction step



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

- Build

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

Distribution-free marginal validity is achievable: cont'd

Conformal Prediction (Vovk et al., 2005; Papadopoulos et al., 2002; Lei et al., 2018) builds an estimated predictive set $\widehat{C}_\alpha$ based on n data points.

Conformal prediction achieves marginal validity (Vovk et al., 2005)

$\widehat{C}_\alpha$ outputted by conformal prediction is such that for any distribution $\mathcal{D}$ on $(\mathcal{X}, \mathcal{Y})$, it holds:

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

✗ Marginal coverage: $\mathbb{P}_{\mathcal{D}^{\otimes(n+1)}}\left(Y^{(n+1)} \in \widehat{C}_\alpha\left(X^{(n+1)}\right) \not{|X^{(n+1)} = x}\right) \geq 1 - \alpha.$

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 $\mathcal{D}$, it holds:

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

Limits of distribution-free conditional predictive uncertainty quantification

Informative conditional coverage as such is impossible

Impossibility results (Vovk, 2012; Lei and Wasserman, 2014)[1]

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:

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

↪ distribution-free X-conditional hardness result apply beyond CP

↪ X-conditional estimators are overly large *even on easy cases*

↪ the lower bound is tight

Naive estimator

$\mathcal{C}_\alpha(\cdot;\xi) \equiv \mathcal{Y}\mathbb{1}\{\xi \leq 1-\alpha\} + \emptyset\mathbb{1}\{\xi > \alpha\}$, where $\xi \sim \mathcal{U}([0,1])$.

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

Getting closer to X-conditional coverage

	X-conditionally valid	Non X-conditionally valid
"Non smooth" distribution	X-cov.: ✓ Length: ✓	X-cov.: ✗ Length: `not relevant`
"Smooth" distribution	X-cov.: ✓ Length: ✗	X-cov.: $\approx$ Length: ✓

- 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)
- Approximate conditional coverage
  $\hookrightarrow$ Romano et al. (2020); Guan (2022); Jung et al. (2023); Gibbs et al. (2023)
  Target $\mathbb{P}(Y^{(n+1)} \in \widehat{C}_\alpha\left(X^{(n+1)}\right) | X^{(n+1)} \in \mathcal{R}(x)) \geq 1 - \alpha$

Non exhaustive references.

Definition of distribution-free group conditional validity ($\mathcal{G}$CV)

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

$\mathcal{G}$ a set of "groups" (i.e., define G a random variable taking its values in $\mathcal{G}$).

Distribution-free $\mathcal{G}$-conditional validity ($\mathcal{G}$CV)

$\widehat{C}_\alpha$ achieves distribution-free $\mathcal{G}$-conditional validity if for any distribution $\mathcal{D}$ on $(\mathcal{X}, \mathcal{G}, \mathcal{Y})$, it holds that:

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

Hardness of distribution-free group conditional coverage

General $\mathcal{G}$CV hardness result (Z., Josse, Romano and Dieuleveut, 2024)[2]

If any $\widehat{C}_\alpha$ is distribution-free $\mathcal{G}$-conditionally valid then **for any distribution** $\mathcal{D}$, for any group $g \in \mathcal{G}$ such that $\mathcal{D}_G(g) > 0$, it holds:

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

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

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

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

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

Sketch of proof

Let $\widehat{C}_\alpha$ be distribution-free $\mathcal{G}$-conditionally valid.

- ▷ Let $\mathcal{D}$ a distribution on $(\mathcal{X}, \mathcal{G}, \mathcal{Y})$.
  - ↪ By assumption, $\widehat{C}_\alpha$ achieves $\mathcal{G}$CV under $\mathcal{D}$.

  Fix a group g such that $\mathcal{D}_G(g) > 0$.
- ▷ Build $\mathcal{Q}$ another distribution such that:
  - $\mathcal{Q}$ equals $\mathcal{D}$ whenever $G \neq g$,
  - $\mathcal{Q}$ is arbitrarily spread on Y (e.g. $\mathcal{U}([-R, R])$ for $R > 0$) when $G = g$.
  - ↪ By assumption, $\widehat{C}_\alpha$ achieves $\mathcal{G}$CV under $\mathcal{Q}$ also.
    Therefore, $\text{mes}\left(\widehat{C}_\alpha\left(X^{(n+1)}, g\right)\right)$ is arbitrarily large under $\mathcal{Q}^{\otimes(n+1)}$ with probability at least $1 - \alpha$.

⇒ By construction, $\mathcal{D}$ and $\mathcal{Q}$ are close in total variation.
They are separated of at most $\mathcal{D}_G(g)$.
  - ↪ Thus, the total variation distance between $\mathcal{D}^{\otimes(n+1)}$ and $\mathcal{Q}^{\otimes(n+1)}$ is of at most $\Delta_{g,n} \leq \mathcal{D}_G(g)\sqrt{n+1}$.

Combining, we get the result under $\mathcal{D}^{\otimes(n+1)}$.

Restricting the link between G and (X or Y) does not allow informative $\mathcal{G}$CV

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

If any $\widehat{C}_\alpha$ is $\mathcal{G}$CV under $G \perp\!\!\!\perp X$, then for any distribution $\mathcal{D}$ such that $G \perp\!\!\!\perp X$, for any group g such that $\mathcal{D}_G(g) > 0$, it holds:

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

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

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

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

$\Rightarrow$ need to restrict both the link between G and X, as well as between G and Y.

Analogous statements are also available for the classification framework.

Implications for $\mathcal{G}$CV in practice

	$\mathcal{G}$-conditionally valid even when $G \not\perp\!\!\!\perp (X, Y)$	$\mathcal{G}$-conditionally valid at most if $G \perp\!\!\!\perp (X, Y)$
"Non smooth" distribution	$\mathcal{G}$-cov.: ✓ Length: ✓	$\mathcal{G}$-cov.: ✗ Length: `not relevant`
"Smooth" distribution	$\mathcal{G}$-cov.: ✓ Length: ✗	$\mathcal{G}$-cov.: $\approx$ Length: ✓

Application to learning with missing covariates

Missing values are ubiquitous and challenging

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

Y	X_1	X_2	X_3
22	5	6	3
19	6	8	NA
19	5	3	6
7	NA	9	NA
13	4	9	0
20	NA	NA	1
9	8	NA	4

Mask $M =$

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

$\hookrightarrow$ 2^d potential masks.

$\hookrightarrow$ M can depend on X or Y (depending on the missing mechanism[3]).

$\Rightarrow$ Statistical and computational challenges.

[3] Three mechanisms connecting X and M from Rubin (1976), *Inference and missing data*, Biometrika

Supervised learning with missing values: impute-then-predict

Impute-then-predict 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 data:

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

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

Goals of predictive uncertainty quantification with missing values

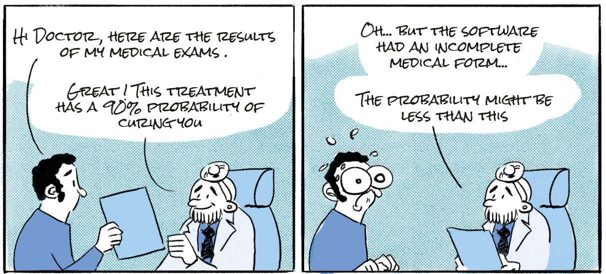
Goal: predict $Y^{(n+1)}$ with **confidence** $1-\alpha$, i.e. build the smallest $\mathcal{C}_\alpha$ such that:

1. Marginal Validity (MV)

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

2. Mask-Conditional-Validity (MCV)

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



Illustrations @theoremlinger

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 ϕ:

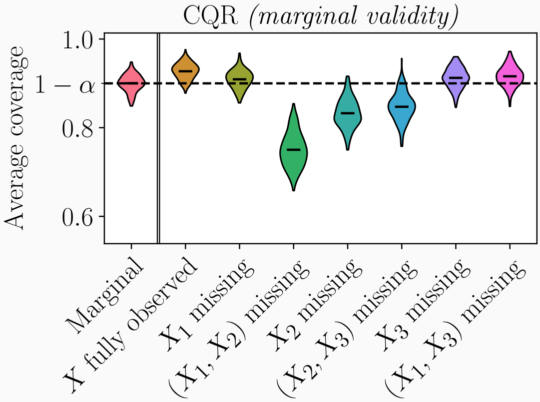
$\left(\phi\left(X^{(k)}_{\text{obs}(M^{(k)})}, M^{(k)}\right), Y^{(k)}\right)_{k=1}^{n}$ are **exchangeable**.

$\Rightarrow$ Conformal Prediction (CP), applied on an imputed data set still enjoys marginal guarantees:

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

CP 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!

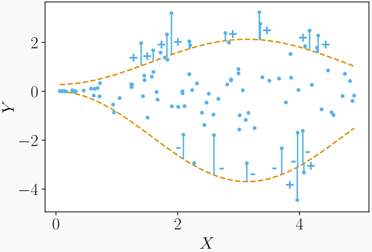
✗ Mask-conditional-validity (MCV) is not attained

↪ Missing values induce heteroskedasticity

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

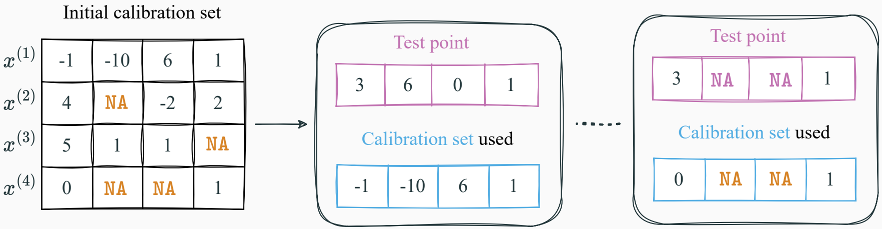
Conformalization step is independent of the important variable: the mask!

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)!



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

Under the assumptions that:

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

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

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

$\hookrightarrow$ **Experiments beyond independence:** under various MAR and MNAR mechanisms, and to some extent when $Y \not\perp\!\!\!\perp M|X$, CP-MDA-Nested* maintains empirical MCV.

Validities 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:

1. Marginal Validity (MV)

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

2. Mask-Conditional-Validity (MCV)

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

	Exisiting approaches	CP-MDA-Nested* (Z., Josse, Romano and Dieuleveut, 2024)
(MV)	✓ (Z., Dieuleveut, Josse, and Romano, 2023)	✓
(MCV)	✗	✓ under $M \perp\!\!\!\perp (X, Y)$

Take-home-messages

- Distribution-free group-conditional-coverage is hard to ensure theoretically on "rare" groups
- Weaker notions are empirically achievable
- These hardness results disappear if $G \perp\!\!\!\perp (X, Y)$
- This strong assumption is relevant in the missing values context
- We propose an algorithm achieving MCV under $G \perp\!\!\!\perp (X, Y)$, empirically robust when $G \not\perp\!\!\!\perp (X, Y)$

Thanks for listening and feel free to reach out to us!

Questions?

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.

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

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.

References iv

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

Zaffran, M., Dieuleveut, A., Josse, J., and Romano, Y. (2023). Conformal prediction with missing values. In *Proceedings of the 40th International Conference on Machine Learning*. PMLR.

Zaffran, M., Josse, J., Romano, Y., and Dieuleveut, A. (2024). Predictive uncertainty quantification with missing covariates. *arXiv:2405.15641*.