Quantmetry
Part of Capgemini Invent

Perspectives & Roundtables

MAPIE General Meeting

2023/11/24

Thibault Cordier
PhD, Data Scientist
tcordier@quantmetry.com

Vincent Blot
PhD Student, Data Scientist
vblot@quantmetry.com

Louis Lacombe
Data Scientist
llacombe@quantmetry.com

Candice Moyet
Data Scientist
cmoyet@quantmetry.com

Nicolas Brunel
Scientific Director
nbrunel@quantmetry.com

Session 1

MAPIE

Introductory session:
Where does MAPIE stand?

15 min

Session 1: Where does MAPIE stand

Subject: present an overview of the MAPIE package (i.e., cartography of features and roadmap)

1) Presentation of MAPIE:

- MAPIE Team
- Missions of MAPIE Team
- History of MAPIE from 2021 to 2023
- Decision tree / Feature matrix of MAPIE

2) Future directions:

- Tentative Roadmap up to 2024 (main focus, priorities, etc.)
- What are your problem with MAPIE?

MAPIE Team & Contributors

MAPIE Team 2024


Thibault Cordier


Vincent Blot


Louis Lacombe


Candice Moyet

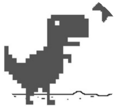

Nicolas Brunel

Emeritus Contributors of MAPIE

Contributors of MAPIE


Vianney Taquet


Sofiane Ziane


JumpingDino


remiadon


alize-papp


AndreaPi


Grégoire Martinon


Thomas Morzadec


Arnaud Capitaine


dan1elherbst


AdirthaBorgohain


cmougan

MAPIE - Model Agnostic Prediction Interval Estimator

scikit-learn-contrib / **MAPIE** Public

Fork 80 Starred 1k

Python library, open source and scikit-learn compatible, for estimating confidence intervals in classification and regression tasks.

- **MAPIE** is an open-source Python library hosted on scikit-learn-contrib project that allows you to:

  1) easily **compute conformal prediction intervals/sets** with controlled marginal coverage rate for regression, classification (binary and multi-class) and time series.

  2) easily **control risks** (such as coverage, recall or any other non-monotone risk) for more complex tasks (multi-label classification, semantic segmentation, …).

  3) easily **wrap any model** (*scikit-learn, tensorflow, pytorch*, ...).

- **MAPIE** is designed and conceived **for academic and industrial uses**.

SPONSORS

Région îledeFrance

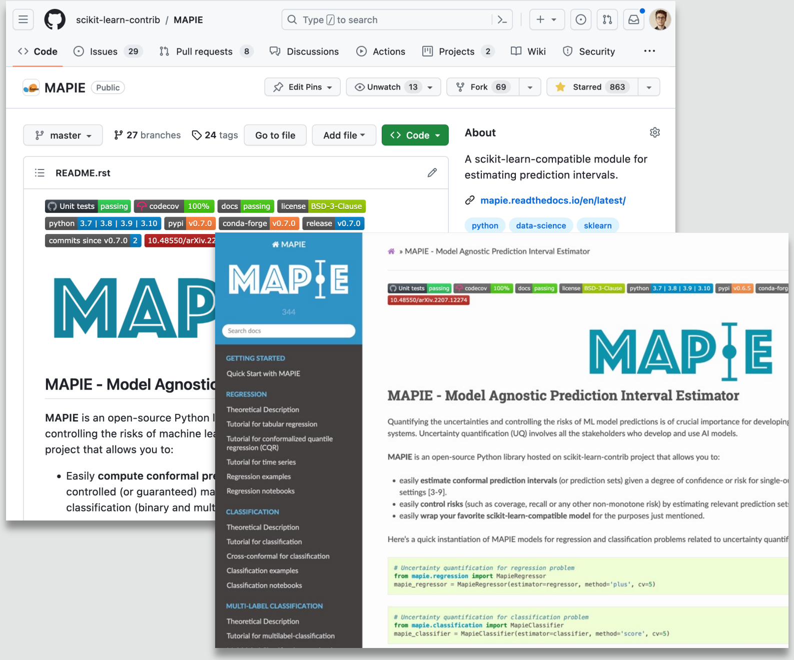
Confiance ai

Quantmetry
Part of Capgemini Invent

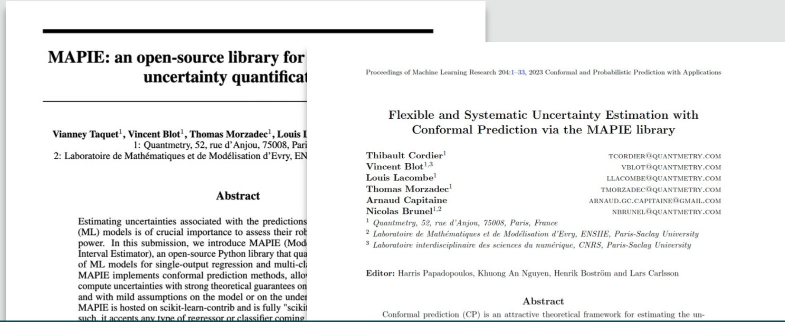
GitHub Read the Docs codecov 100% scikit learn contrib python Package Index CONDA-FORGE

Missions of MAPIE Team

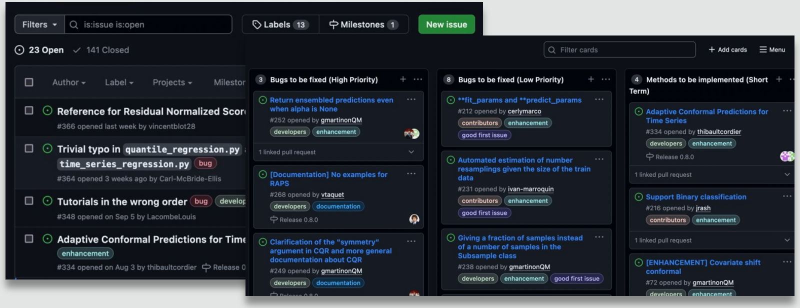
Github Project



Scientific Publications



Code Maintenance



History of MAPIE - from 2021 to 2023

0.1.0 (2021-04-27)

0.2.0 (2021-05-21)

0.3.0 (2021-09-10)

0.4.0 (2022-06-24)

0.5.0 (2022-10-20)

0.6.0 (2023-01-19)

0.7.0 (2023-09-14)

2021 – Start of MAPIE

- **MapieRegressor**: split/cross-conformal methods (Jackknife+ , CV+, Jackknife+-after-Bootstrap)
- **MapieClassifier**: split-conformal method, LAC and APS methods
- **Metrics:** marginal coverage and mean width scores for regression and classification

2022 - Consolidation

- **MapieQuantileRegressor:** CQR
- **MapieTimeSeriesRegressor:** EnbPI
- **MapieClassifier:** cross-conformal method, Top-K and RAPS methods
- **ConformityScore:** Absolute and Gamma conformity scores, (un)signed conformity scores

2023 – Continuity

- **MapieMultiLabelClassifier:** RCPS and CRC methods, Lean-Then-Test framework (false discovery rate control)
- **MapieCalibrator**: Top-Label method
- **Statistical calibration tests**: Kolmogorov-Smirnov, Kuiper and Spiegelhalter tests
- **ConformityScore:** residual normalised conformity score
- **Metrics**: HSIC, Size-Stratified Coverage, ECE and Top-label ECE scores

Trust AI

What can you find in the release 0.7.0 of MAPIE?

Summary table of algorithms implemented in MAPIE

Release 0.7.0

Task	Feature	Algorithm	Reference
PI/PS	MapieRegressor MapieClassifier	**Jackknife/CV+**	Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. "Predictive inference with the jackknife+." Ann. Statist., 49(1):486–507, (2021).
		Jackknife/CV+ ab	Kim, Byol, Chen Xu, and Rina Barber. "Predictive inference is free with the jackknife+-after-bootstrap." Advances in NeurIPS 33 (2020): 4138-4149.
Prediction intervals (PI)	AbsoluteConformityScore GammaConformityScore ResidualNormalizedScore	**Absolute Score**	Vovk, Vladimir, Alexander Gammerman, and Glenn Shafer. Algorithmic Learning in a Random World. Springer Nature, 2005
		Gamma Score	Cordier, Thibault, Vincent Blot, Louis Lacombe, Thomas Morzadec, Arnaud Capitaine, Nicolas Brunel "Flexible and Systematic Uncertainty Estimation with Conformal Prediction via the MAPIE library", COPA (2023)
		Normalised Score	Papadopoulos, Harris, Proedrou, Kostas, Vovk, Volodya, and Gammerman, Alex. "Inductive confidence machines for regression". In Machine Learning: ECML (2002).
	MapieTimeSeriesRegressor	**EnbPI**	Xu, Chen, and Yao Xie. "Conformal prediction interval for dynamic time-series." International Conference on Machine Learning. PMLR, (2021).
	MapieQuantileRegressor	**CQR**	Romano, Yaniv, Evan Patterson, and Emmanuel Candes. "Conformalized quantile regression." Advances in neural information processing systems 32 (2019).
Prediction sets (PS)	MapieClassifier	**LAC / LABEL**	Sadinle, Mauricio, Jing Lei, and Larry Wasserman. "Least ambiguous set-valued classifiers with bounded error levels." Journal of the American Statistical Association 114.525 (2019): 223-234.
		APS	Romano, Yaniv, Matteo Sesia, and Emmanuel Candès. "Classification with valid and adaptive coverage." Advances in NeurIPS 33 (2020): 3581-3591.
		Top-K	Angelopoulos, Anastasios, et al. "Uncertainty sets for image classifiers using conformal prediction." International Conference on Learning Representations (2021).
		RAPS	Angelopoulos, Anastasios, et al. "Uncertainty sets for image classifiers using conformal prediction." International Conference on Learning Representations (2021).
Control Risks (CR)	MapieMultiLabelClassifier	**RCPS**	Bates, Stephen, et al. "Distribution-free, risk-controlling prediction sets." Journal of the ACM (JACM) 68.6 (2021): 1-34.
		CRC	Angelopoulos, Anastasios N., Stephen, Bates, Adam, Fisch, Lihua, Lei, and Tal, Schuster. "Conformal Risk Control." (2022).
		LTT	Angelopoulos, Anastasios N., Stephen, Bates, Emmanuel J. Candès, et al. "Learn Then Test: Calibrating Predictive Algorithms to Achieve Risk Control." (2022).
Calib.	MapieCalibrator	**Top-label**	Gupta, Chirag, and Aaditya K. Ramdas. "Top-label calibration and multiclass-to-binary reductions." arXiv preprint arXiv:2107.08353 (2021).

Decision tree / Feature matrix of MAPIE

If I have a predictive model, MAPIE can give me guarantees and insights on the quality of the predictions.

1. Compute uncertainty intervals / sets:
“I want to ensure that the true labels are covered."

Task		Split	Cross	Metrics
Regression	-			Mc/Ma
	Time Series			Mc
Classification	Binary	See “3. Calibrate my model” (1)		
	Multiclass			Mc/Ma
	Multilabel	See “2. Control a risk” (2)		
Object Detection	Binary	*Calib.: OK*		Mc/Ma
	Multiclass	*Class.: OK*		Mc/Ma
Instance Segmentation		*Class.: OK*		Mc/Ma

2. Control the risk of my model: “I want to guarantee that my risk is under control with a probability guarantee.”

Task	Risk	Split	Cross	Metrics
Multilabel Classification	with Precision and Recall Guarantees (2)			Mc
Selective Regression	with MSE Control			-
	with OOD Detection			
Selective Classification	with Accuracy Control			-
	with OOD Detection			
Selective Generation	with Auxiliary Control			-
Object Detection	with Coverage, and Recall Guarantees			-
Instance Segmentation	with mIOU Guarantee			-

3. Calibrate my model:
“I want to ensure that scores given by my model are probabilities”

Task		Split	Cross	Metrics
Classification	Binary (1)			Mc/H
	Multiclass			Mc
	Multilabel			-

4. Test a hypothesis: “I want to make sure that assumptions on my data are satisfied or detect deviation of the model.”

Test	Split	Cross	Metrics
Calibration test of binary classifier			H
Exchangeability test (distribution drift)			-
Performance stability test (perf. drift)			-
Anomaly detection			-

Colour code: Not available | Available in MAPIE

Legend

Colour code:

Alternative solution	No literature	Not available in MAPIE
Non-adaptive approach	Adaptive approach	Available on scikit-learn

Split: Split data into training / calibration sets (split-conformal method)
Cross: Does not impact training data (cross-conformal method)
Mc: Coverage metrics available
Ma: Adaptability metrics available
H: Hypothesis Testing

Tentative Roadmap up to 2024

0.8.0 (2024-x-x)

Main Focus - P1

- [Session 2]: **Mondrian Conformal Prediction:** validity within categories thanks to Mondrian approach for both regression and classification
- [Session 2]: **Adaptive Conformal Prediction:** approaches based on non-conformity scores distribution estimation
- [Session 3]: **Hypothesis Testing:** toolbox of hypothesis tests for exchangeability and performance drift
- **Risk control:** selective regression and classification with LTT

Priorities - P2

- **Interoperability**: support natively pytorch, tensorflow, transformers, etc.
- **Large Scale Deployment**: accelerate code, be robust to memory problems...
- **Use Cases**: instance segmentation, object detection, generation (tabular data, text), etc.

To be completed...

- Other propositions, with your contributions and your comments!

Schedule

Session 1: Overview of MAPIE — 15 min

Session 2: Adaptive Uncertainty Quantification — 15 min + 10 min of Q&A

Session 3: Hypothesis Testing — 10 min + 10 min of Q&A

Session 4: Round Table — 15 min

Teams Chat:

Q&A at the end of each session

Session 2

MAPIE

Session 2:
Adaptive Uncertainty Quantification

15 min

Session 2: Adaptive Uncertainty Quantification

1) Presentation:

- What is Adaptive Uncertainty Quantification? Why?
- What already exists in MAPIE

2) Future directions:

- Methods for conditional coverage
- Methods based on estimated scores distribution
- Metrics for adaptive uncertainty

3) Opening: round table

What is Adaptive Uncertainty Quantification? Why?

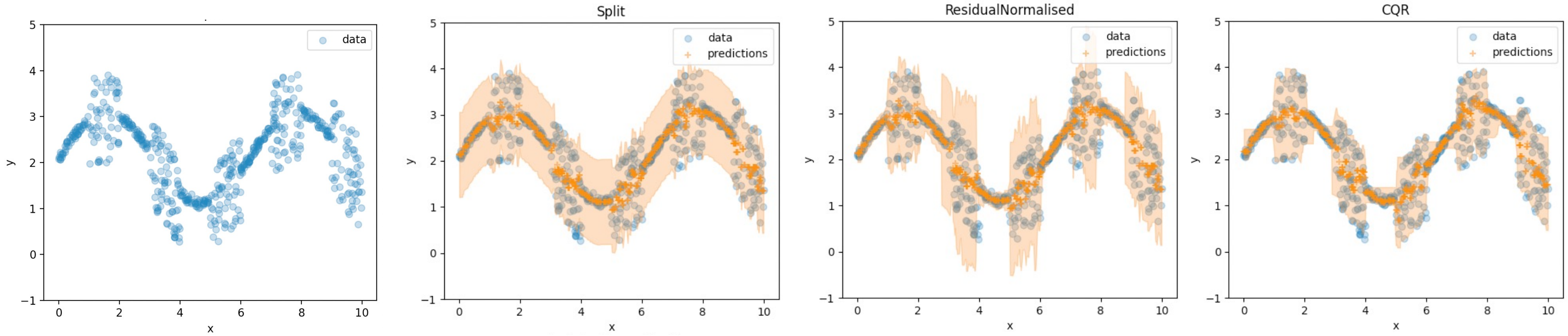
The ideal framework for uncertainty quantification (UQ):

1) Be distribution-free
2) Be valid in finite samples
3) **Satisfy the conditional coverage guarantee:**

- $\mathbb{P}\left(Y_{n+1} \in \hat{C}(X_{n+1}) \middle| X_{n+1} = \mathrm{x}\right) = 1 - \alpha$

What is a good adaptive method for UQ?

1. Predictions sets should be of **various sizes.**
2. **Size of the prediction sets ⇔ Uncertainty of the model.**
3. **Prediction coverage should be guaranteed locally** (for any prediction), and not just globally (on average).



Issues and limitations:

- **Conditional coverage** is proven to be **impossible to satisfy** ([1-3]) without assumptions on the distribution or algorithm.
- **Marginal coverage** is easily obtained with conformal prediction but without local interpretation: $\mathbb{P}(Y_{n+1} \in \hat{C}(X_{n+1})) = 1 - \alpha$

[1] Foygel Barber, R., Candes, E. J., Ramdas, A., & Tibshirani, R. J. (2021). The limits of distribution-free conditional predictive inference. Information and Inference: A Journal of the IMA, 10(2), 455-482.
[2] Vladimir Vovk. (2012) Conditional validity of inductive conformal predictors. In Asian conference on machine learning, pages 475–490.
[3] Vovk, V., Gammerman, A., & Shafer, G. (2005). Algorithmic learning in a random world (Vol. 29). New York: Springer.

What already exists in MAPIE for adaptive uncertainty quantification

Our MAPIE methods are mainly based on **split-CP** with **calibration dataset** $D_n^{Cal} = \{(x_1, y_1), \dots (x_n, y_n)\}$ and NCS $\{s_i = s(x_i, y_i)\}_{1 \le i \le n}$.
They compute the **quantiles of the NCS** $Q_{n,\alpha}^- / Q_{n,\alpha}^+$ to estimate the **bounds of the prediction intervals** $\hat{C}_{n,\alpha}(X_{n+1})$ given a test data X_{n+1}.

Regression Methods in MAPIE

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

$$\hat{C}_{n,\alpha}(X_{n+1}) = \left[\hat{\boldsymbol{q}}_{\alpha_{low}}(\boldsymbol{X}_{n+1}) - Q_{n,\alpha}^-, \hat{\boldsymbol{q}}_{\alpha_{up}}(\boldsymbol{X}_{n+1}) + Q_{n,\alpha}^+\right]$$
$$s(x, y) = \max(y - \hat{\boldsymbol{q}}_{\alpha_{low}}(\boldsymbol{x}), \hat{\boldsymbol{q}}_{\alpha_{up}}(\boldsymbol{x}) - y)$$

- **Residual Normalized Score,** *Lei et al. 2016*

$$\hat{C}_{n,\alpha}(X_{n+1}) = [\hat{\mu}(X_{n+1}) - Q_{n,\alpha}^- * \hat{\boldsymbol{\sigma}}(\boldsymbol{X}_{n+1}), \hat{\mu}(X_{n+1}) + Q_{n,\alpha}^+ * \hat{\boldsymbol{\sigma}}(\boldsymbol{X}_{n+1})]$$
$$s(x, y) = \frac{|y - \hat{\mu}(x)|}{|\hat{\boldsymbol{\sigma}}(\boldsymbol{x})|}$$

- **Gamma Score,** *Cordier et al. 2023*

$$\hat{C}_{n,\alpha}(X_{n+1}) = [\hat{\mu}(X_{n+1}) * (\boldsymbol{1} - \boldsymbol{Q}_{n,\alpha}^-), \hat{\mu}(X_{n+1}) * (\boldsymbol{1} + \boldsymbol{Q}_{n,\alpha}^+)]$$
$$s(x, y) = \frac{|y - \hat{\mu}(x)|}{|\hat{\boldsymbol{\mu}}(\boldsymbol{x})|}$$

Classification Methods in MAPIE

- **Adaptive Prediction Sets (APS),** *Romano et al. 2020*

$$\hat{C}_{n,\alpha}(X_{n+1}) = \{\pi_1, \dots, \pi_k\} \quad where \quad k = \inf\{s(x_n, k) \ge Q_{n,\alpha}\}$$
$$\boldsymbol{s}(\boldsymbol{x}_i, \boldsymbol{k}) = \sum_{j=1}^{k} \hat{\boldsymbol{\mu}}(\boldsymbol{x}_i)_{\pi_j} \quad where \quad \forall\, j \ge k \quad \hat{\mu}(x_i)_{\pi_j} \ge \hat{\mu}(x_i)_{\pi_k}$$

- **Regularized APS (RAPS),** *Angelopoulos et al. 2020*

$$\hat{C}_{n,\alpha}(X_{n+1}) = \{\pi_1, \dots, \pi_k\} \quad where \quad k = \inf\{s(x_n, k) \ge Q_{n,\alpha}\}$$
$$\boldsymbol{s}(\boldsymbol{x}_i, \boldsymbol{k}) = \sum_{j=1}^{k} \hat{\boldsymbol{\mu}}(\boldsymbol{x}_i)_{\pi_j} + \boldsymbol{\lambda}(\boldsymbol{k} - \boldsymbol{k}_{reg})^+ \quad where \quad \forall\, j \ge k \quad \hat{\mu}(x_i)_{\pi_j} \ge \hat{\mu}(x_i)_{\pi_k}$$

Issues and limitations:

- These methods requires **an auxiliar model** (calculation costs, depend on their performance, must be trained externally, on other calibration data, ...).
- They are based only **on marginal coverage guarantee** in their design.

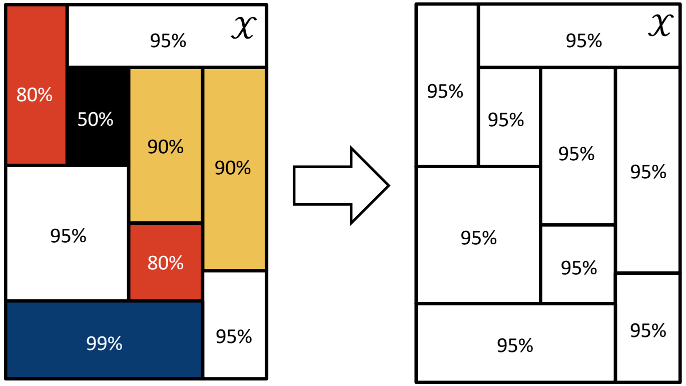
Methods for conditional coverage

Goal: **Group-conditional coverage** or $\mathbb{P}\left(Y_{n+1} \in \hat{C}(X_{n+1}) \middle| Z_{n+1} = z\right) = 1 - \alpha$ when Z_{n+1} is a categorical variable.

Mondrian Conformal Predictors

Algorithmic Learning in a Random World, *Vovk et al. 2005(v1), 2022(v2)* [1]

- Partition of $\mathcal{X}$ $\longrightarrow$ "conformalization" on each part of the partition



Split-CP Mondrian-CP

Ideas / Openings

- **Mondrian-CP Process:** $\forall\, \mathcal{P}_i \in \mathcal{P}(\mathcal{X})$,
  1. Define α_i a risk and $n_i = |\mathcal{P}_i|$
  2. Estimate $\mathcal{S}_i$ the NCS of the observations in $\mathcal{P}_i$
  3. Estimate Q_{n_i,α_i} the quantile $(1 - \alpha_i)$ of $\mathcal{S}_i$
- **Equivalent to applying several Split-CPs to each part.**

- **Generalizing the conditioning criterion:**
  - Based on an exogenous criterion
  - Based on a partition of the $\mathcal{X}$ space
  - Based on a partition of the $\mathcal{Y}$ space
  - Based on a partition of the $\mathcal{X}x\mathcal{Y}$ space

- **Integration perspectives:**
  - Applicable to regression and classification tasks
  - Needs to be adapted for flexible integration and use

[1] Vovk, V., Gammerman, A., & Shafer, G. (2005/2022). Algorithmic learning in a random world (Vol. 29). New York: Springer.

Methods based on estimated scores distribution (without covariate shift)

Goal: Generalizing group-conditional coverage with **a proxy / a relaxation of conditional coverage.**

Proxy by estimating conditioned NCS distribution

Adaptive Conformal Prediction by Reweighting Nonconformity Scores, *Amoukou et al. 2023* [1]

- The prediction set is built as follows:

$$\boldsymbol{C_{n,\alpha}(X_{n+1}) = \left\{ y \mid s(X_{n+1}, y) \leq \hat{Q}\left(1 - \tilde{\alpha};\ \hat{F}_S(\cdot \mid X_{n+1} = x)\right)\right\}}$$

where $\hat{F}_S(s|X_{n+1} = x) = \sum_{i=1}^{n} w(x_i, x)\, \mathbf{1}\{s_i \leq s\}$ is the conditional f.d.r. estimated by a Random Forest, and $\tilde{\alpha}$ is selected for reaching target marginal coverage $1 - \alpha$.

- Conditional-training can be obtained with additional correction, and asymptotic conditional coverage.

- Extension of **Localized conformal prediction: a generalized inference framework for conformal prediction,** *Guan 2022* [2] based on Nadaraya-Watson estimators.

Interpolation between marginal and conditional coverage overlapping groups

Conformal Prediction with Conditional Guarantees, *Gibbs et al. 2023* [3]

- Generalization of Mondrian to overlapping groups by relaxing conditional coverage:

$$\mathbb{P}(Y_{n+1} \in \hat{C}(X_{n+1}) \mid X_{n+1} = x) = 1 - \alpha, \quad \text{for all } x$$

$$\Longleftrightarrow$$

$$\mathbb{E}[f(X_{n+1})(\mathbf{1}\{Y_{n+1} \in \hat{C}(X_{n+1})\} - (1 - \alpha))] = 0, \quad \text{for all measurable } f.$$

- The prediction set is built as follows:

$$\boldsymbol{C_{n,\alpha}(X_{n+1}) = \left\{ y \mid s(X_{n+1}, y) \leq \hat{g}_{S(X_{n+1}, y)}(X_{n+1}) \right\}}$$

by replacing the constant quantile of prediction scores $\{s_i = s(x_i, y_i)\}$ by a quantile regression $g(\cdot) \in \mathcal{F}$ (a RKHS) on X_{n+1} (with pinball loss):

- Possibility of control for finite (e.g. partition of sets gives back Mondrian) and infinite dimensional function space $\mathcal{F}$.

[1] Amoukou, S. I., & Brunel, N. J. (2023). Adaptive Conformal Prediction by Reweighting Nonconformity Score. arXiv preprint arXiv:2303.12695.
[2] Guan, L. (2023). Localized conformal prediction: A generalized inference framework for conformal prediction. Biometrika, 110(1), 33-50.
[3] Gibbs, I., Cherian, J. J., & Candès, E. J. (2023). Conformal Prediction With Conditional Guarantees. arXiv preprint arXiv:2305.12616.

What metrics already exist in MAPIE for adaptive uncertainty

Why: To find out whether the model is **uniformly good at being adaptive** to model uncertainty

How: Global measure expressing local coverage (without depending on X or other parameters)

Existing in MAPIE

- **Size-Stratified Coverage (SSC),** *Angelopoulos et al. 2021* [1]

$$\mathrm{SSC}(X, Y; \mathcal{G}) = \min_{g \in \mathcal{G}} \frac{1}{|\mathcal{I}_g|} \sum_{i \in \mathcal{I}_g} \mathbb{1}_{\{Y_i \in C(X_i)\}}$$

- **Hilbert-Schmidt Independence Criterion (HSIC),** *usage proposed by Feldman et al. 2021* [2]

Correlation measure between the coverage (X) and the interval size ($\ell(X)$). By considering two separable RKHS on X and $\ell(X)$, HSIC is defined as the Hilbert Schmidt norm of the cross-covariance operator.

$$\mathrm{HSIC}(X, \ell(X); \mathcal{F}, \mathcal{G}) = \|\mathrm{C}_{X\ell(X)}\|_{HS}^2$$

- **[SOON] Coverage Width-based Criterion (CWC),** *usage proposed by Jensen et al. 2022* [3]

Trade-off between the prediction interval normalized average width (PINAW) and the prediction interval coverage probability (PICP).

$$\mathrm{CWC}(\eta) = (1 - \mathrm{PINAW}) e^{-\eta(\mathrm{PICP} - (1-\alpha))^2}$$

Metrics related to Mondrian-CP

- **Group-Conditional Coverage:**

Given a partition of the observations with respect to a categorical criterion (sub-groups) defined by an exogenous rule given by user, compute the **coverage within the categories**:

$$\forall z \in \mathcal{Z},$$
$$\mathrm{cov}(z) = \mathbb{E}_{X,Y}\left[\, \mathbb{1}_{\{Y \in C(X)\}} \,\middle|\, Z = z \,\right]$$

Issues and limitations:

- These measures are **not easy to use and interpret**; they require arbitrary binarization, a mixing parameter or depend on substitution variables.

[1] Angelopoulos, A. N., & Bates, S. (2023). Conformal prediction: A gentle introduction. Foundations and Trends® in Machine Learning, 16(4), 494-591.
[2] Feldman, S., Bates, S., & Romano, Y. (2021). Improving conditional coverage via orthogonal quantile regression. Advances in neural information processing systems, 34, 2060-2071.
[3] Jensen, V., Bianchi, F. M., & Anfinsen, S. N. (2022). Ensemble conformalized quantile regression for probabilistic time series forecasting. IEEE Transactions on Neural Networks and Learning Systems.

Metrics for adaptive uncertainty

Why: To find out whether the model is **uniformly good at being adaptive** to model uncertainty

How: Global measure expressing local coverage (without depending on X or other parameters)

Ideas / Openings

- **Deficit and Excess, *Seedat et al. 2023* [1]**

Deficit: interval shortfall, when the true value y lies outside the predicted interval.

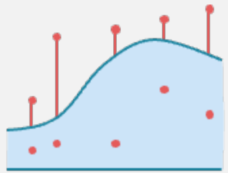
- measures the **under-coverage** of the prediction sets.

Excess: additional width included not needed to capture the true value y

- measures the **over-coverage** of the prediction sets.

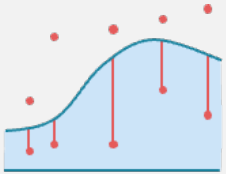
Deficit:

$$\mathbb{E}[\mathbb{1}_{Y \notin [l(X),u(X)]} \cdot min\{|Y - l(X)|, |Y - u(X)|\}]$$



Excess:

$$\mathbb{E}[\mathbb{1}_{Y \in [l(X),u(X)]} \cdot min\{|Y - l(X)|, |Y - u(X)|\}]$$



- **Correlation between interval size and residuals**

Beyond measuring adaptability

Statistical inference for fairness auditing, *Cherian et al. 2023* [2]

- Construct a statistical certificate for controlling the disparity of a performance metric $L(\hat{\mu}(X), Y)$ between a group G and the global target of the model:

$$\underbrace{\epsilon(G)}_{\text{disparity}} := \underbrace{\mathbb{E}_P[L(f(X),Y) \mid (X,Y) \in G]}_{\text{group-specific}} - \underbrace{\theta_P}_{\text{target}} .$$

- Bootstrap is used for computing lower bound ϵ_{lb} such that:

$$\lim_{n\to\infty} \mathbb{P}\left(\epsilon_{lb} \leq \epsilon(G)\ \forall\ G\right) \geq 1 - \alpha$$

- and we can test $H_0(G)$: $\epsilon(G) \leq \epsilon$, with adapted threshold and FWER gives:

$$\lim_{n\to\infty} \mathbb{P}\left(\exists\ G \text{ falsely certified}\right) \leq \alpha$$

[1] Seedat, N., Jeffares, A., Imrie, F., & van der Schaar, M. (2023, April). Improving adaptive conformal prediction using self-supervised learning. In International Conference on Artificial Intelligence and Statistics (pp. 10160-10177). PMLR.
[2] Cherian, J. J., & Candès, E. J. (2023). Statistical Inference for Fairness Auditing. arXiv preprint arXiv:2305.03712.

Round table

10 min

Session 3

MAPIE

Session 3:
Hypothesis Testing

10 min

Session 3: Hypothesis Testing

1) Presentation:

- Calibration tests in MAPIE
- Why hypothesis testing?
- Two families of hypothesis testing

2) Future directions:

- Hypothesis testing for detecting significant performance drift

3) Opening: round table

- Which hypothesis tests do you need?

What already exists in MAPIE for hypothesis testing?

Even if the calibration of binary classifier is not implemented in MAPIE as it already exists in scikit-learn, one would like to test if the model is calibrated or not:

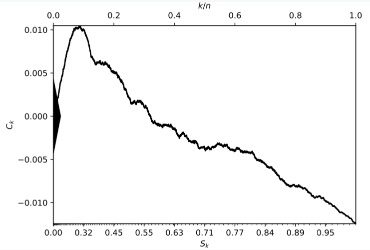
[H0] My model is well calibrated vs. [H1] My model is not calibrated

Cumulative differences

$$C_k = \frac{1}{N}\sum_{i=1}^{N}(y_i - s_i)$$

Characteristic length

$$\sigma = \frac{1}{N}\sqrt{\sum_{i=1}^{N} s_i(1 - s_i)}$$



Kolmogoriv-Smirnov test [1, 2, 3]

- **Test statistic:**

$$\boldsymbol{G} = \max_{1 \le k \le N} |\boldsymbol{C_k}|$$

G/σ converges in distribution to the maximum absolute value of 1D Brownian motion

p-value: [2, 3] propose closed-form formulas for the cumulative distribution function (CDF) of $\boldsymbol{G/\sigma}$ is given by :

$$\boldsymbol{p = 1 - CDF(G/\sigma)}$$

Kuiper test [1, 2, 3]

- **Test statistic:**

$$\boldsymbol{H} = \max_{1 \le k \le N} |\boldsymbol{C_k}| - \min_{1 \le k \le N} |\boldsymbol{C_k}|$$

H/σ converges in distribution to the maximum absolute value of 1D Brownian motion

p-value: [2, 3] propose closed-form formulas for the cumulative distribution function (CDF) of $\boldsymbol{H/\sigma}$ is given by :

$$\boldsymbol{p = 1 - CDF(H/\sigma)}$$

Spiedgelhalter test [4]

- **Test statistic:**

$$B = \frac{1}{N}\sum_{i=1}^{N}(y_i - s_i)(1 - 2s_i) + \frac{1}{N}\sum_{i=1}^{N} s_i(1 - s_i)$$

$$Z = \frac{B - \mathbb{E}[B]}{\sqrt{Var[B]}} = \frac{\sum_{i=1}^{N}(y_i - s_i)(1 - 2s_i)}{\sqrt{\sum_{i=1}^{N}(1 - 2s_i)^2 s_i(1 - s_i)}}$$

p-value: This statistic follows a normal distribution of cumulative distribution CDF, so that we state the associated p-value:

$$\boldsymbol{p = 1 - CDF(Z)}$$

[1] Arrieta-Ibarra I, Gujral P, Tannen J, Tygert M, Xu C. Metrics of calibration for probabilistic predictions. The Journal of Machine Learning Research. 2022 Jan 1;23(1):15886-940.
[2] Tygert M. Calibration of P-values for calibration and for deviation of a subpopulation from the full population. arXiv preprint arXiv:2202.00100. 2022 Jan 31.
[3] D. A. Darling. A. J. F. Siegert. The First Passage Problem for a Continuous Markov Process. Ann. Math. Statist. 24 (4) 624 - 639, December, 1953.
[4] Spiegelhalter DJ. Probabilistic prediction in patient management and clinical trials. Statistics in medicine. 1986 Sep;5(5):421-33.

Why hypothesis testing?

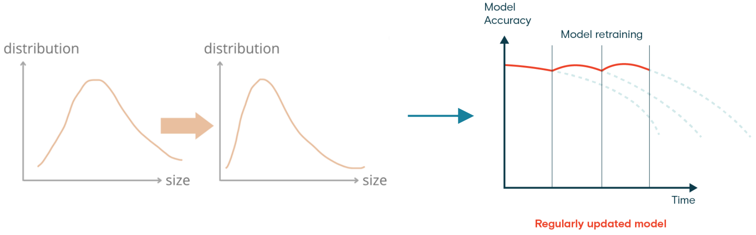
Preventive and **on-line ways of monitoring** the strength of evidence against the **assumption of exchangeability.**

Theoretical use for CP (Before using MAPIE)

- Can I use the conformal prediction framework?

- **Hypothesis testing (upstream)**
  - **[H0] exchangeability vs [H1] non exchangeability**

Business application (When the model is deployed)



- How do I detect changes in the data?
- Do I need to retain my model?

- **Performance testing (downstream)**
  - **[H0] stable performance vs [H1] derived performance**

Two families of hypothesis testing

Distribution drift / exchangeability testing

Algorithmic Learning in a Random World, *Vovk et al. 2022* [1]

- **Part III – Testing Randomness**
  - *"**Conformal testing** is a way of testing the IID assumption based on conformal prediction."* [2]
  - *"Valid testing procedures are equated with **test martingales**"* [2]
  - Usual Testing (batch) vs. Conformal Testing (online)
- Overview of uses suggested by Vovk
  - Testing Exchangeability
  - Testing for Concept and Label Shift
  - When to retrain: CUSUM, Shiryaev-Roberts, Variable & Fixed Training Schedules

Testing randomness *by Vladimir Vovk 2020* [2]

Testing exchangeability: fork-convexity, supermartingales, and e-processes, *by Aaditya Ramdas et al. 2021* [3]

- **High-potential subject to be explored**
  - Call for proposals for tests relevant to MAPIE users

Performance drift testing

Bounded metrics:
Regression: coverage
Classification: coverage, precision & recall

Tracking the risk of a deployed model and detecting harmful distribution shifts *by Aleksandr Podkopaev, Aaditya Ramdas 2021* [4]

Unbounded metrics:
Regression: MSE, SSR

Tips for moving from an unbounded metric to a bounded metric or to other directions.

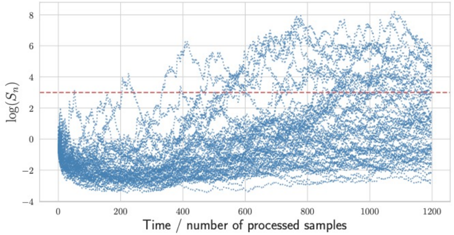
[1] Vovk, V., Gammerman, A., & Shafer, G. (2005, 2022). Algorithmic learning in a random world (Vol. 29). New York: Springer.
[2] Vovk, V. (2021). Testing randomness online. Statistical Science, 36(4), 595-611.
[3] Ramdas, A., Ruf, J., Larsson, M., & Koolen, W. M. (2022). Testing exchangeability: Fork-convexity, supermartingales and e-processes. International Journal of Approximate Reasoning, 141, 83-109..
[4] Podkopaev, A., & Ramdas, A. (2021). Tracking the risk of a deployed model and detecting harmful distribution shifts. arXiv preprint arXiv:2110.06177.

Hypothesis testing for detecting significant performance drift

Motivation: Use a metric to control for **harmful** distribution drifts that have a **business** impact to **re-train models** at a given time.

Issue

In a classification setting, the harmful shit only occurs once the argmax changes value. In this scenario, we see that even though the marginal probability class of class 1 goes from 0.1 to 0.45 in increments of 0.05, test martingales would raise an error multiple times.



Issues and limitations:

- The data needs to be **IID or independent,** not exchangeable.
- **Risk needs to be upper bounded (or lower)**, hence, for classification: precision, recall and for regression: coverage → not only business metrics.

Hypothesis Testing in a business setting

Tracking the risk of a deployed model and detecting harmful distribution shifts, *Aleksandr Podkopaev and Aaditya Ramdas, 2021*

- $l(\cdot,\cdot)$: the loss function, chosen to be monitored
- $f: X \to Y$: the predictors
- $R(f) = \mathbb{E}[l(f(X), Y)]$: expected loss; called the risk of f
- $\widehat{U}_S(f)$: upper confidence bound on the source risk
- $\hat{L}_T^{(t)}(f)$: lower confidence bound on the target risk continuously updated for the target risk as new data points are observed at time t
- $H_0 = R_t(f) \leq R_S(f) + \varepsilon_{tol}$

Algorithm 1 Sequential testing for an absolute increase in the risk.

Input: Predictor f, loss ℓ, tolerance level ε_{tol}, sample from the source $\{(X_i, Y_i)\}_{i=1}^{n_S}$.

```
procedure
    Compute the upper confidence bound on the source risk $\widehat{U}_S(f)$;
    for $t = 1, 2, \ldots$ do
        Compute the lower confidence bound on the target risk $\widehat{L}_T^{(t)}(f)$;
        if $\widehat{L}_T^{(t)}(f) > \widehat{U}_S(f) + \varepsilon_{\text{tol}}$ then
            Reject $H_0$ (equation 1) and fire off a warning.
```

Quantmetry
Part of Capgemini Invent

Round table

10 min

Session 4

MAPIE

Session 4:
Round Table / Retex

15 min

MAPIE

Thank you for your attention.