A Glimpse on Conformal prediction

Michel Lutz

Institut für
Medizinische Systembiologie

Agenda

1. **What is CP?**
   - Coverage guarantee
   - LAC, APS
   - Split, CV+, Jackknife+
2. **Experiment**
   - Data
   - Results
3. **Where to start?**
4. **History of CP**
5. **Marginal vs. Conditional Coverage**
   - Group-balanced CP
   - Class-balanced CP
6. **Conformal Change Point Detection**
   - CTM
   - PERMAD Dataset

Heuristic uncertainty → Conformal prediction → Rigourus uncertainty

point prediction → prediction set
with coverage guarantee

prediction interval
without coverage guarantee → prediction interval
with coverage guarantee

uncalibrated probabilistic distribution → well calibrated probabilistic distribution
(Venn predictors)

Coverage guarantee

$$1-\alpha \quad \leq \quad \mathbb{P}\Big(\mathbf{y}_{test} \in \mathbf{C}(\mathbf{x}_{test})\Big) \quad \leq \quad 1-\alpha+\frac{1}{n+1}$$

- **model agnostic**
- **distribution free**
- **finite sample size**
- **minimal assumptions**
  - (exchangeability)

Exchangeability

- Weaker then iid.

$$(X_1, Y_1), \ldots, (X_i, Y_i), \ldots, (X_n, Y_n), (X_{n+1}, Y_{n+1})$$

swap

$$(X_1, Y_1), \ldots, (X_i, Y_i), \ldots, (X_n, Y_n), (X_{n+1}, Y_{n+1})$$

- After swapping the datasets cannot be distinguished

Full Conformal Prediction

- train one model for each label in the label space

$$\forall y \in \mathcal{Y} \quad (X_1, Y_1), \dots, (X_N, Y_N), (X_{n+1}, y) \longrightarrow f^y$$

$$(X_{n+1}, Y_{true})$$ exchangeable to all other points

Conformity scores

$$s_i^y = S\Big((X_i, Y_i), f^y\Big) \quad \text{,for } i = 1, \dots, n$$

$$s_{n+1}^y = S\Big((X_{n+1}, y), f^y\Big)$$

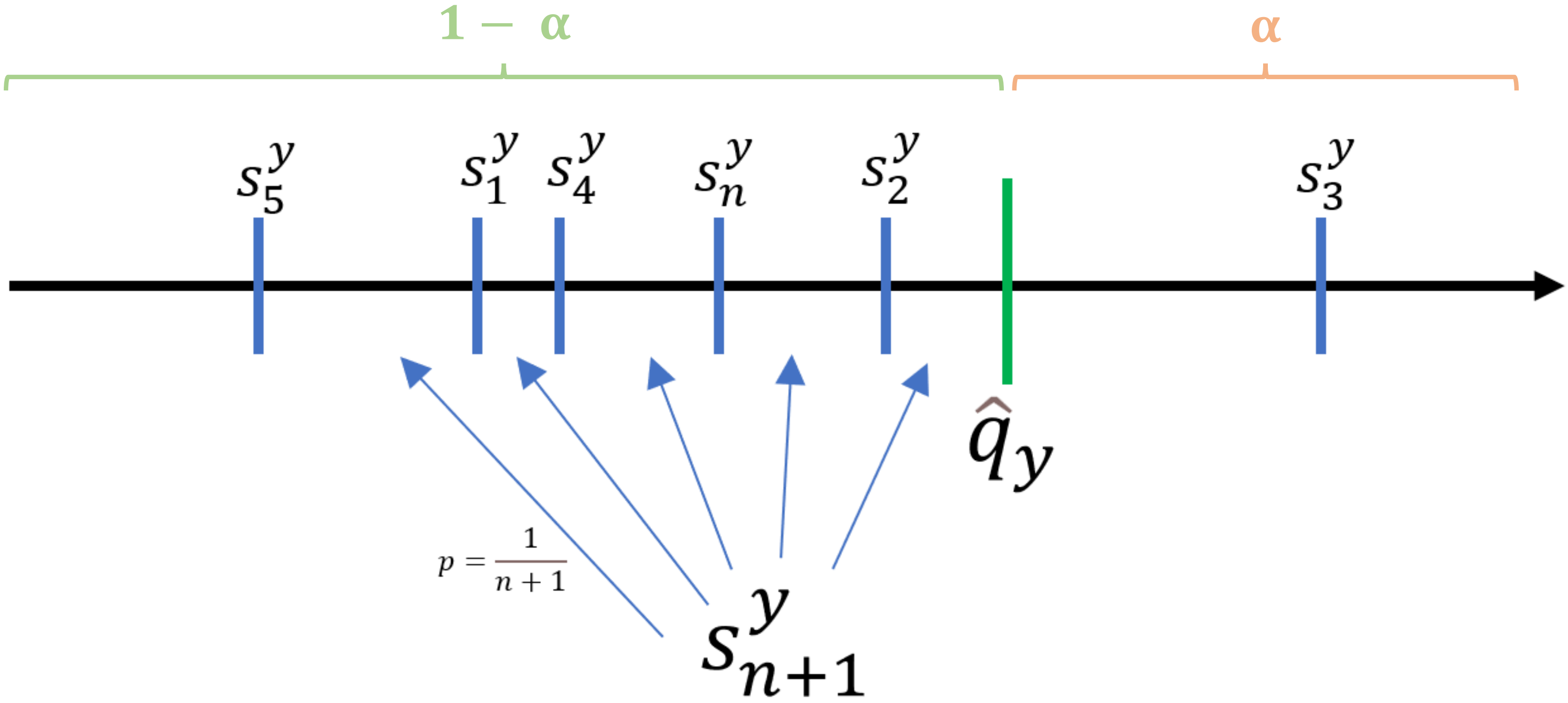
1-alpha quantile

$$\hat{q}^y = \text{Quantile}\left(s_1^y, \dots, s_n^y; \frac{\lceil (1-\alpha)(n+1) \rceil}{n}\right)$$

Prediction set

$$\mathbf{C}(X_{test}) = \left\{ y : s_{n+1}^y \leq \hat{q}^y \right\}$$

1-alpha quantile



Reason: exchangeability of the data

Conformity score

$$s_i = S\Big((X_i, Y_i), \hat{f}\Big)$$

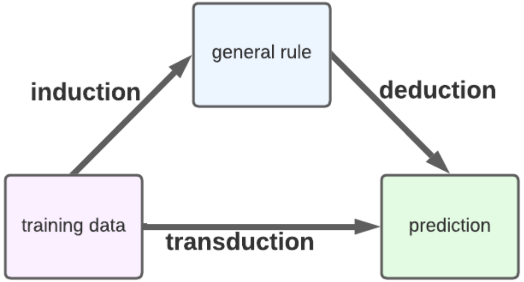
- The higher, the more strange the data point is for the rest of the data
- Can be an arbitrary function
- Coverage guarantee holds for all conformity scores
- BUT! : Informativeness (size of the prediction set) depends on S
- S = Noise function -> Prediction set has size 1-alpha of the label space

Split vs. Full Conformal prediction

Full CP

- transductive
- train one model for each possible label
- start for each prediction from scratch

\+ hight data efficiency
\- high computational effort



Split CP

- inductive
- split data in training
- and calibration set
- only train model once

\- low data efficiency
\+ low computational effort
\+ can be used for pretrained models

Split conformal prediction

Calibration set $$Z_{\text{calib}} = (X_1, Y_1), \ldots, (X_n, Y_n)$$

Conformity score $$s_i = S\Big((X_i, Y_i), \hat{f}\Big)$$

1-alpha Quantile $$\hat{q} := \frac{\lceil (1-\alpha)(n+1) \rceil}{n} \text{ Quantile of: } s_1, \ldots, s_n$$

Prediction set $$\mathbf{C}(X_{test}) = \Big\{ y : s(X_{test}, y) \leq \hat{q} \Big\}$$

[2] Angelopoulos, A. N., & Bates, S. (2021)

Conformal recipe (for split CP)

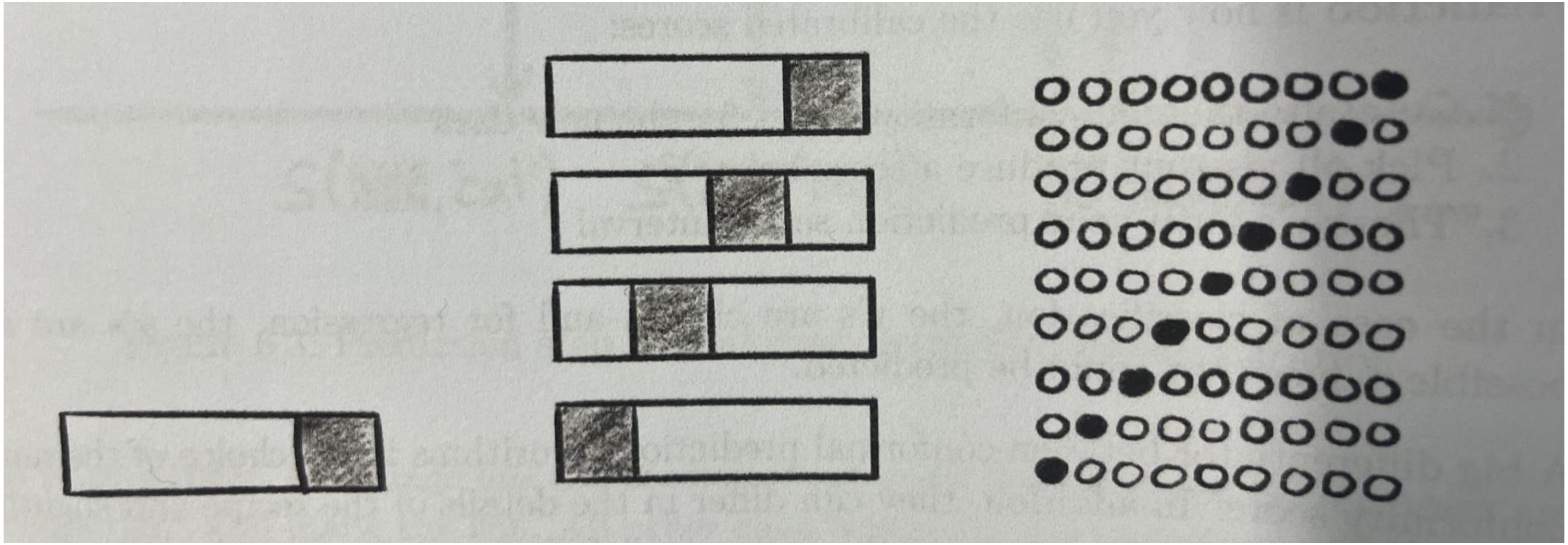
1. Identify a heuristic notion of uncertainty provided by $\hat{f}$
2. Define a score function $S\left(X_i, Y_i; \hat{f}\right)$ based on the heuristic notion of uncertainty.
3. Compute $\hat{q}$ as the $\frac{\lceil(1-\alpha(n+1)\rceil}{n}$ quantile of the calibration scores

$$s_1 = S\left(X_1, Y_1; \hat{f}\right), \ldots, s_n = S\left((X_n, Y_n; \hat{f}\right)$$

4. Calculate the prediction sets for a new data point X_{test} as:

$$\mathbf{C}(X_{test}) = \left\{y : S(X_{test}, y; \hat{f}) \leq \hat{q}\right\}$$

Split, CV+, Jackknife+



Split CV+ Jackknife+

Least Ambiguous set-valued Classifier (lac)

$$s_i = 1 - \hat{f}_{y=y_i}(x_i)$$

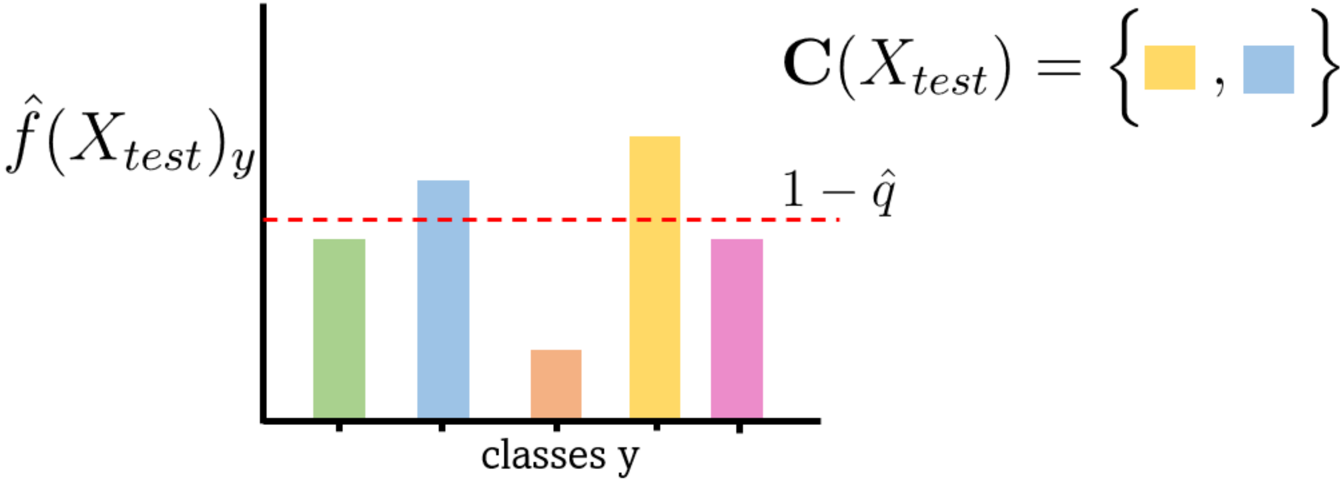
$$\mathbf{C}(X_{test}) = \left\{ y : \hat{f}(X_{test})_y \geq 1 - \hat{q} \right\}$$

- uses only the probability of the true label
- smallest prediction sets (on average)
- lacks adaptivity

[4] Lei, J., & Wasserman, L. (2014)

Least Ambiguous set-valued Classifier (lac)

$$\mathbf{C}(X_{test}) = \left\{ y : \hat{f}(X_{test})_y \geq 1 - \hat{q} \right\}$$



Adaptive prediction scores

$\pi(x)$ Permutation that sorts classes from most to least likely

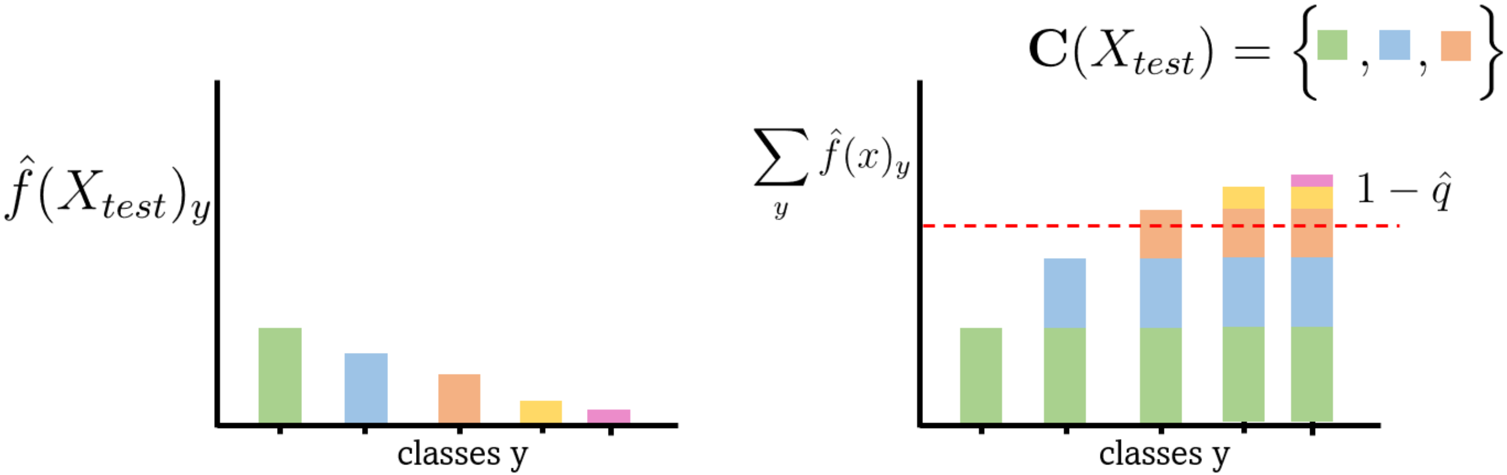
$$s(x, y) = \sum_{j=1}^{c} \hat{f}(x)_{\pi_j(x)} \ , \ \text{where } y = \pi_c(x)$$

- includes the difficulty of a prediction point
- utilizes the scores of all classes, not just the true class
- **more adaptive**

[5] Lei, J. (2014)

Adaptive prediction scores

$$s(x,y)=\sum_{j=1}^{c}\hat{f}(x)_{\pi_j(x)}\ ,\ \text{where } y=\pi_c(x)$$



Adaptive classification with split-conformal calibration (aps)

generalized conditional quantile function for an arbitrary $\tau \in [0,1]$

$$L(x; f, \tau) = \min\{c \in 1, ..., C \ : \ f_{(1)}(x) + f_1(x) + ... + f_c(x) \geq \tau\}$$

$$S(x, u; f, \tau) = \begin{cases} \text{corresponding } y \text{ for the } L(x; f, 1-\alpha) - 1 \text{ largest } f_y(x), & \text{if } u \geq V(x; f, \tau) \\ \text{corresponding } y \text{ for the } L(x; f, 1-\alpha) \text{ largest } f_y(x), & \text{otherwise} \end{cases}$$

$$E(x, y, u; \hat{f}) = \min\{\tau \in [0,1] : y \in S(x, u; f, \tau)\}$$

- works in principle like the other variant
- includes theoretical guarantees and tie-breaking

[3] Romano, Y., Sesia, M., & Candes, E. (2020)

Adaptive classification with split conformal calibration

Algorithm 1 Adaptive classification with split-conformal calibration

1: **Input:** data $\{X_i, Y_i\}_{i=1}^n$, X_{test} , model $\hat{f}$, α
2: X_{train}, $X_{calib} \leftarrow$ train_test_split($\{X_i, Y_i\}_{i=1}^n$)
3: Train $\hat{f}$ on X_{train}
4: Compute $E_i = E(x_i, y_i, u_i; \hat{f})$ for each $x_i, y_i \in X_{calib}$ with function 11
5: Compute $\hat{Q}_{1-\alpha}(\{E_i\}_{i \in X_{calib}})$ as the $\lceil(1-\alpha)(1-|X_{calib}|\rceil$th largest value in E_i
6: **Output** the prediction set:

$$C_{n,\alpha}^{\mathrm{SC}}(x_{test}) = S(x_{test}, u_{test}; \hat{f}, \hat{Q}_{1-\alpha}(\{E_i\}_{i \in X_{calib}}))$$

using the score function S defined in 8.

[3] Romano, Y., Sesia, M., & Candes, E. (2020)

Adaptive classification with CV+

Algorithm 2 Adaptive classification with CV+ calibration

1: **Input:** data $\{X_i, Y_i\}_{i=1}^{n}$, X_{test} , model $\hat{f}$, number of splits $K \leq n$, α
2: Split data into k random distinct subsets $\mathcal{I}_1, \mathcal{I}_2, ..., \mathcal{I}_k$
3: **for** $k \in \{1, ..., k\}$:
4: Train $\hat{f}^{k(i)}$ on $\{X_i, Y_i\}_{i \in \{1,...,n\} \setminus \mathcal{I}_k}$
5: **Output** the prediction set:

$$C_{n,\alpha}^{\mathrm{CV+}}(x_{n+1}) = \Bigg\{ y \in \mathcal{Y} :$$

$$\sum_{n=1}^{n} \mathbf{1}\left[E(x_i, y_i, u_i; \hat{f}^{k(i)}) \leq E(x_{n+1}, y_{n+1}, u_{n+1}; \hat{f}^{k(i)})\right] \leq \lceil (1-\alpha)(1-|n|) \rceil \Bigg\}$$

where $k(i) \in \{1, .., k\}$ denotes the fold containing the ith sample and using the function E defined in 11.

[3] Romano, Y., Sesia, M., & Candes, E. (2020)

Coverage guarantee for CV+

CV+

$$\mathbb{P}\left[Y_{test} \in C_{n,\alpha}^{\mathrm{CV+}}(x_{test})\right] \geq 1 - 2\alpha - \min\left\{\frac{2(1 - 1\backslash K}{n\backslash K + 1}, \frac{1 - K\backslash n}{K + 1}\right\}$$

Jackknife+

Special case of CV+ with k=n

$$\mathbb{P}\left[Y_{test} \in C_{n,\alpha}^{\mathrm{CV+}}(x_{test})\right] \geq 1 - 2\alpha$$

[3] Romano, Y., Sesia, M., & Candes, E. (2020)

Influence of Calibration Set Size

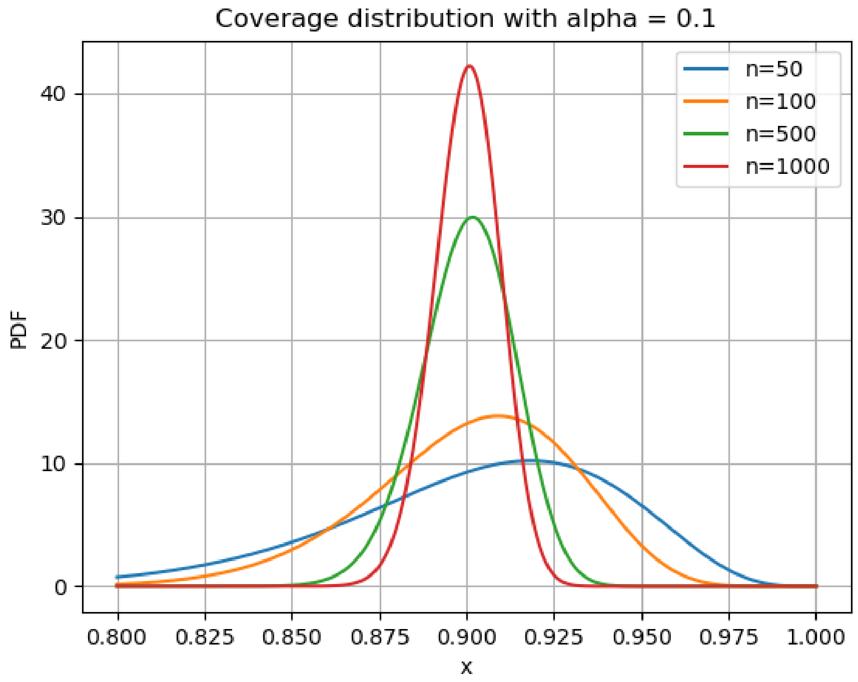
Coverage guarantee holds for coverage of $1 - \alpha$ on **average over the randomness** in the calibration set

ϵ	0.1	0.05	0.01	0.001
$n(\epsilon)$	22	102	9812	244390

Required calibration set size $n(\epsilon)$

for coverage of $1 - 0.9 \pm \epsilon$

with probability $\delta = 0.1$



$$\mathbb{P}\Big(Y_{test} \in C(X_{test}) \mid \{(X_i, Y_i)\}_{i=1}^{y}\Big) \sim Beta(n+1-l, l)\ ,\ l = \lfloor (n+1)\alpha \rfloor$$

[1] Vovk, V. (2012)

Artificial dataset

- 20 features
- 5 classes
- 10.000 datapoints

- 5000 for training/calibration
- (rest) for testing

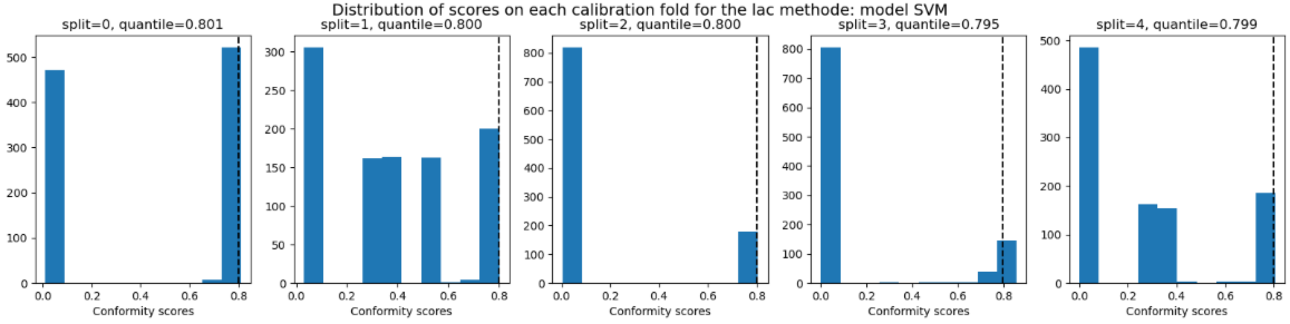
```
def create_artificial_data(n_features, n_classes, sample_size=10000, random_stat=42):
    X, Y = make_classification(
        n_samples=sample_size,
        n_features=n_features,
        n_informative=15,
        n_redundant=2,
        n_repeated=0,
        n_classes=n_classes,
        n_clusters_per_class=1,
        weights=None,
        flip_y=0.001,
        class_sep=1.0,
        hypercube=True,
        shift=0.0,
        scale=2.0,
        shuffle=True,
        random_state=random_stat)
```

Dry Beans dataset

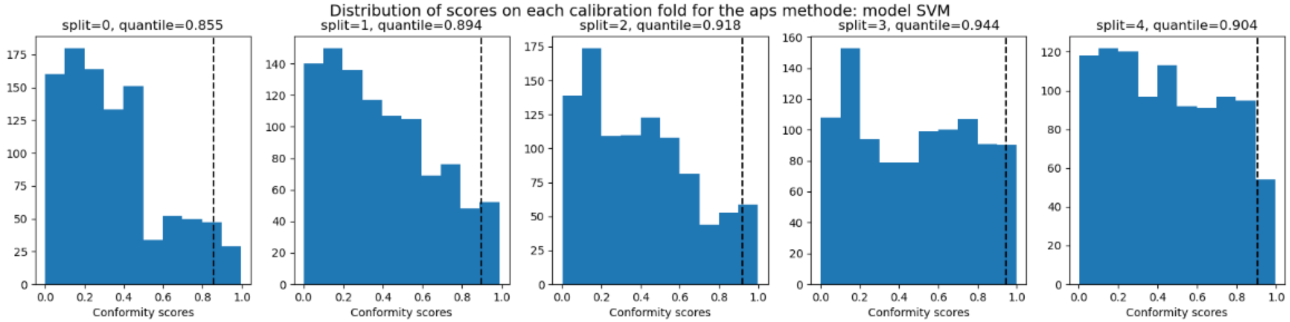
- 13,611 dry beans
- 7 variants (classes)
- 8 features (length, roundness…)

[6] Koklu, M., & Ozkan, I. A. (2020)

Influence of different splits on the conformal scores



(a) least ambiguous set-valued classifier score



(b) adaptive prediction set score

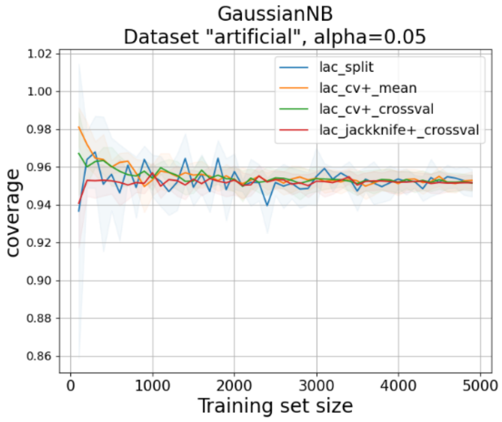
Influence of small datasets on different conformal methods

- **Split** :
  - 20% of the training data as calibration set

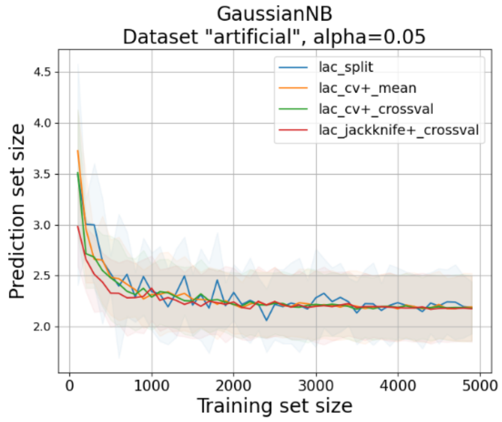
- **CV+** :
  - k=5 splits
  - mean aggregation
  - cross-validation aggregation

- **Jackknife+**
  - like CV+ with k=n

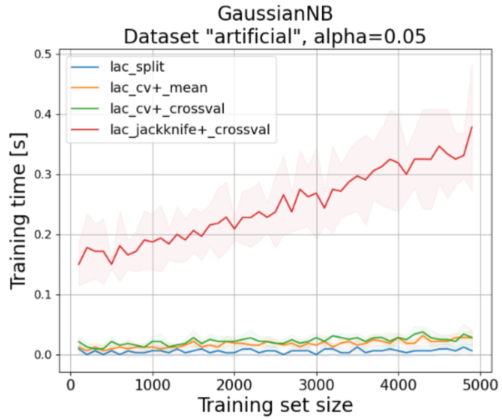
Influence of small datasets on different conformal methods



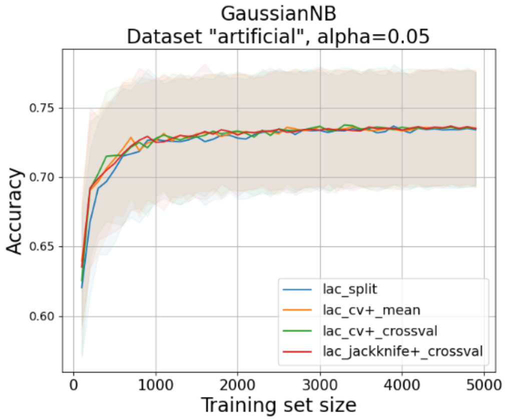
(a) Coverage



(b) Average prediction set size

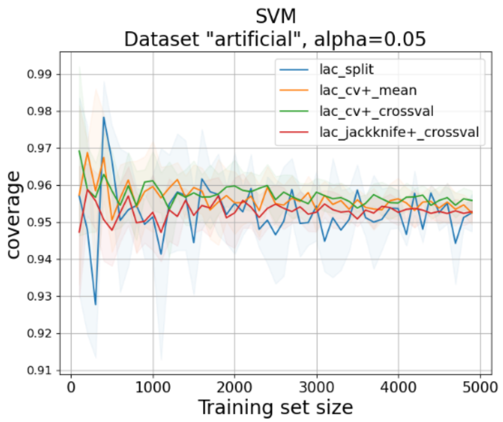


(c) Training time

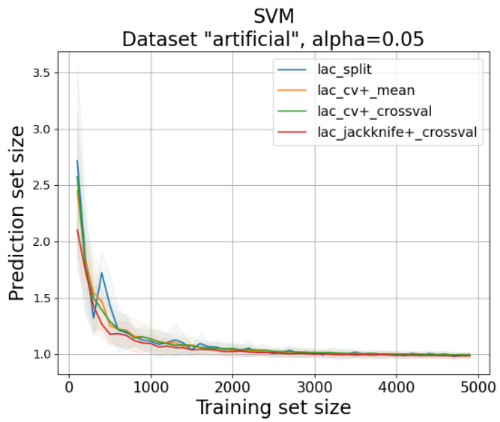


(d) Accuracy

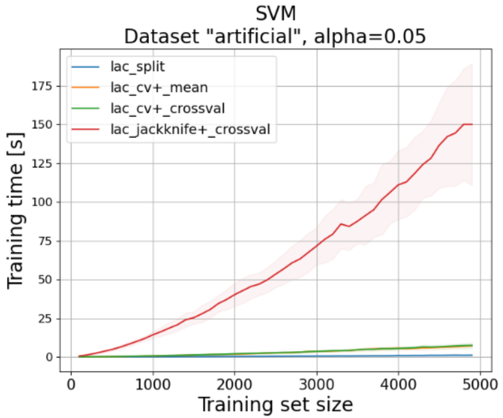
Influence of small datasets on different conformal methods



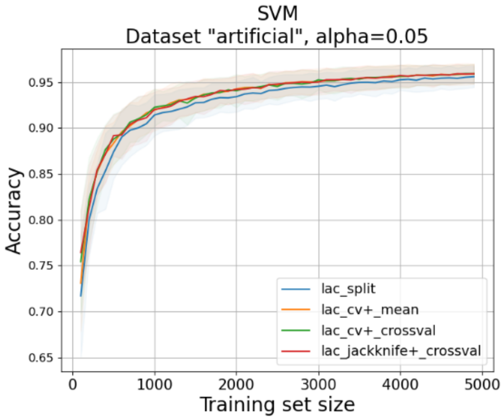
(a) Coverage



(b) Average prediction set size

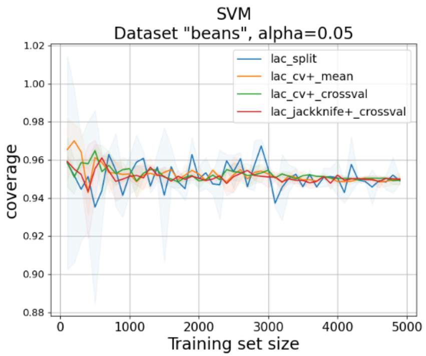


(c) Training time

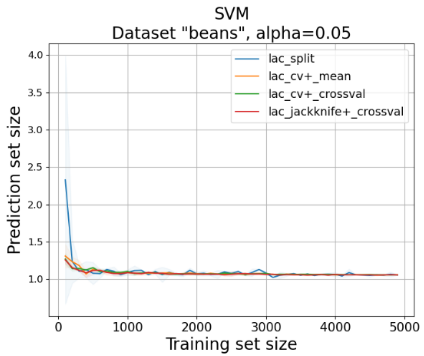


(d) Accuracy

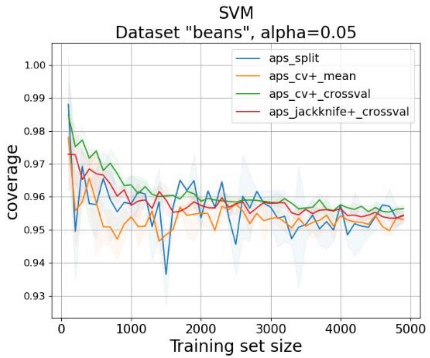
Influence of small datasets on different conformal methods



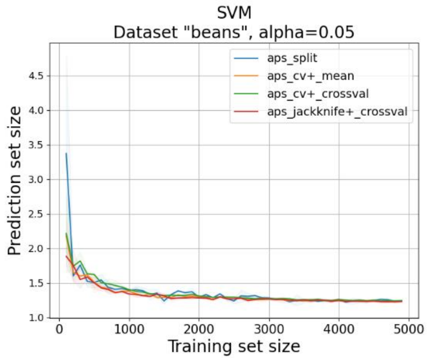
(a) Coverage



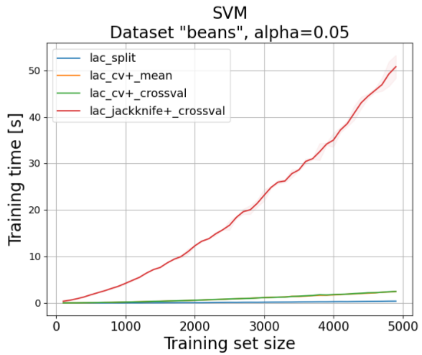
(b) Average prediction set size



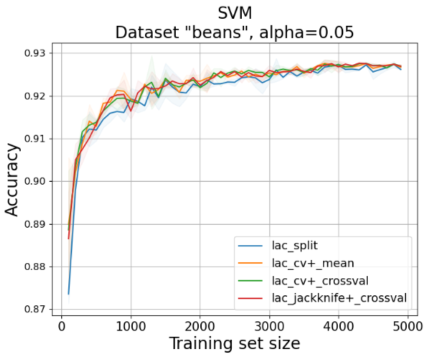
(a) Coverage



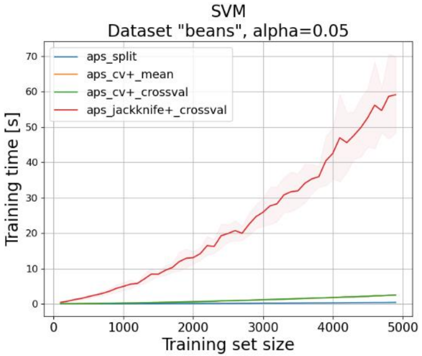
(b) Average prediction set size



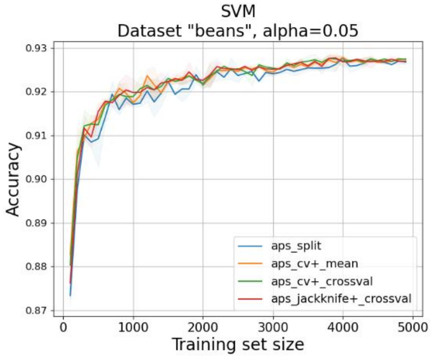
(c) Training time



(d) Accuracy



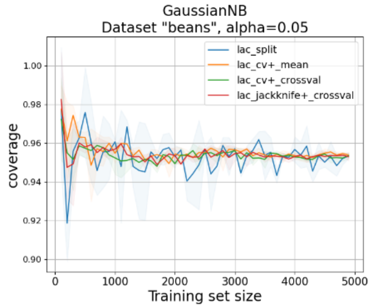
(c) Training time



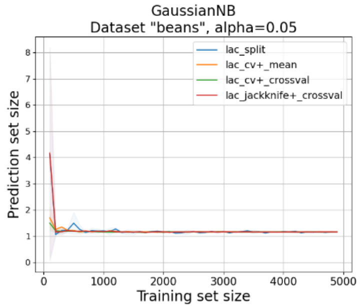
(d) Accuracy

Strange behaviour for aps scores

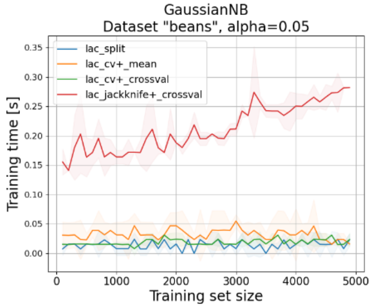
lac



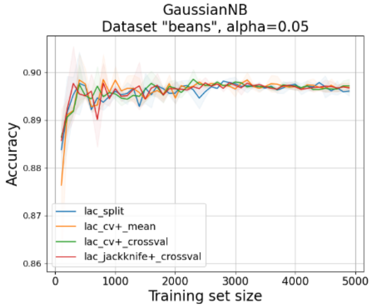
(a) Coverage



(b) Average prediction set size

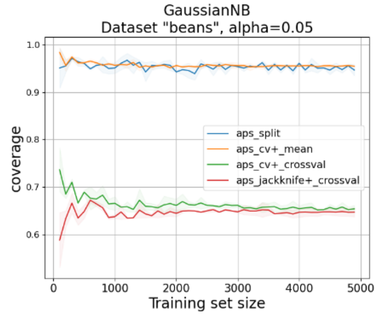


(c) Training time

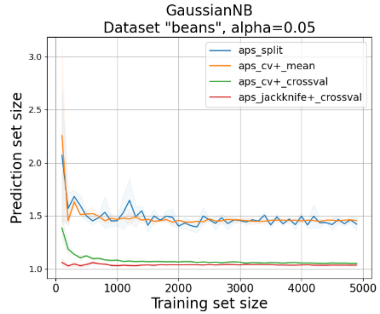


(d) Accuracy

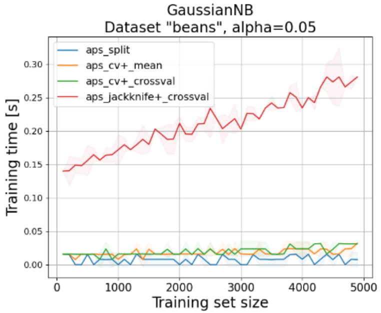
aps



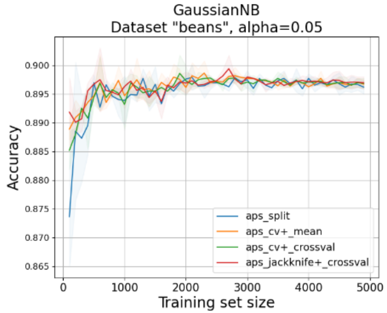
(a) Coverage



(b) Average prediction set size



(c) Training time



(d) Accuracy

Drawback of Conformal Classifiers

- we must be careful when interpreting conformal classifiers
- make exactly α errors on the long run
- an error is when the correct label is not in the prediction set
- the coverage guarantee only applies apriori, once we have seen a specific prediction, we cannot say that the probability for that prediction to be wrong is α

- Example: two class problem
- Prediction sets containing both classes makes no error
- All errors in singletons
- After observing singleton -> probability of an error is much higher then α

Where to start ?

- **“A gentle introduction to conformal prediction and distribution-free uncertainty quantification”**
  - basic overview and foundations
  - [2] Angelopoulos & Bates
  - video tutorial: Gentle Introduction – Tutorial

- **Awesome Conformal Prediction Git Repo**
  - Valery Manokhin
  - newest stuff
  - Tonnes of papers, tutorials, videos, theses,

- **“Conformal prediction: a unified review of theory and new challenges”**
  - [7] Fontana, Matteo, Gianluca Zeni, and Simone Vantini
  - Contains more the continental approach (after Vovk)

- **Project: A Glimpse on Conformal Prediction**
  - Michel Lutz
  - Conformal-prediction-project-repo

END

Appendix:

- History of CP
- Marginal vs. Conditional Coverage

History of Conformal Prediction - I

1960-1980 Andrei Kolmogorov

- Moscow State University
- randomness, complexity and probability
- algorithmically random sequences, finite Bernulli sequences.
- Vladimir Vovk becomes his

1988 PhD Thesis Vovk

- ‘Predictability of algorithmically random’
- role of finite-sample exchangeability in prediction problems

1996-1999 Vovk, Gammerman, Vapnik

- Royal Holloway University of London
- Develop the Conformal Prediction framework

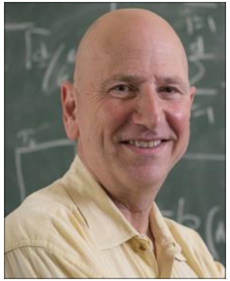
History of Conformal Prediction - III

2014 Sweden

- Ulf Johansson, Henrik Boström & Henrik Linusson
- Jönköping University
- Random forests
- Plenty papers & tutorials
- Venn predictors

2014 Larry Wassermann

- Carnegie Mellon University
- "Ambassadors" for CP in the US
- Class-balanced CP
- Distribution-free predictive inference in regression



2019 Emmanuel Candes

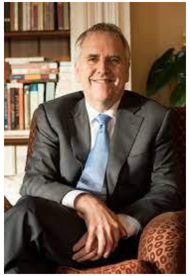
- Stanford University
- Conformalized Quantile Regression
- Plenty papers and fundamental work

History of Conformal Prediction - IV

2020 Adaptive prediction sets
- Romano et al. (Candes group)

2020 Washington Post
- Used CP method to for the U.S. presidential election
- Candes

2021 Michael Jordan
- Anastasios N. Angelopoulos and Stephen Bates
- UC Berkeley
- Large Scale deep learning CP



2021 Conformal Risk Control
- Angelopoulos & Bates

2021 Conformal Outlier Detection
- Bates

History of Conformal Prediction - V

2021 Change point detection
- Vovk

2021 Trak at ICML2021&2022
- Emmanuel Candes

2022 NeurIPS2022
- Emmanuel Candes

2021 Awesome Conformal Prediction
- Git Repo
- Valery Manokhin (PhD Student Vovk)

2022 MAPIE
- Scikit-learn compatible package

2022 CP Beyond Exchangeability
- Candes, Ryan Tibshirani

Marginal vs. conditional coverage

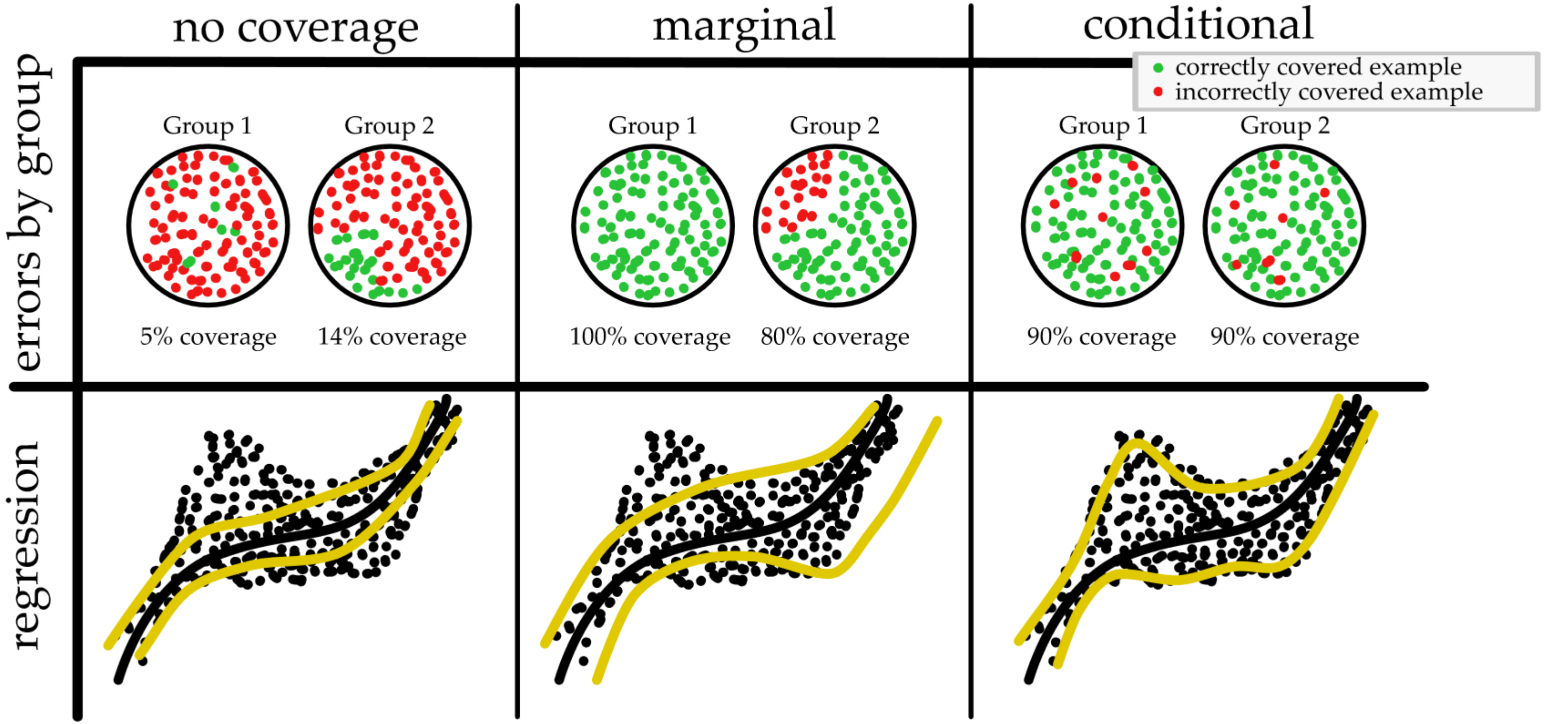
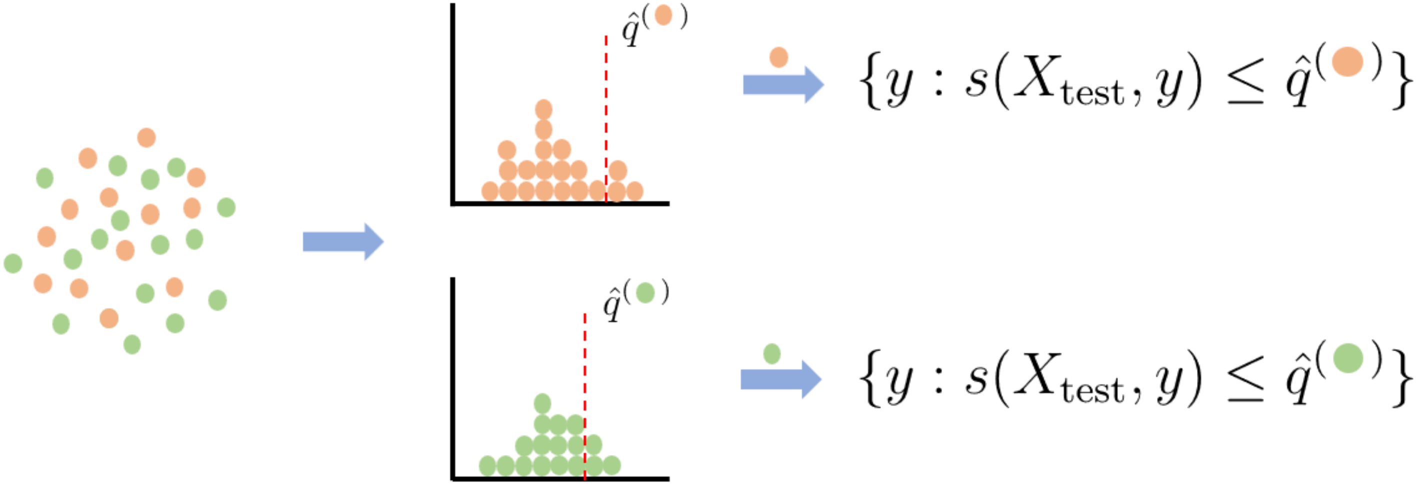


Figure taken from: [2] Angelopoulos, A. N., & Bates

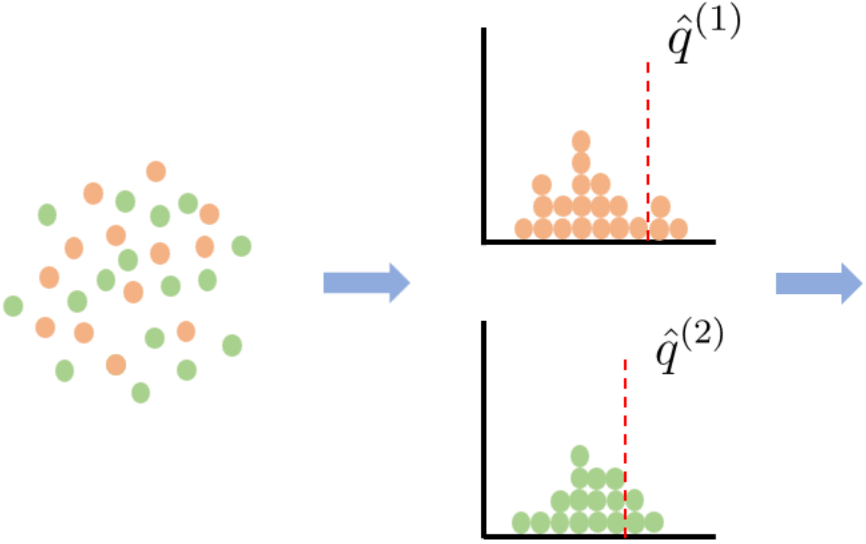
Group-balanced conditional coverage

$$1 - \alpha \leq \mathbb{P}\Big(\mathbf{y}_{test} \in \mathbf{C}(\mathbf{X}_{test}) \mid X_{test} = g_i\Big) \ : \ \forall g \in G$$



Class-balanced conditional coverage

$$1 - \alpha \quad \leq \quad \mathbb{P}\Big(\mathbf{y}_{test} \in \mathbf{C}(\mathbf{X}_{test}) \mid Y_{test} = y\Big)$$



$$\{y : s(X_{\text{test}}, y) \leq \hat{q}^{(\text{y})}\}$$

Additional Experiments – very small datasets

- smallest possible calibration-set size for conformal methods

$$\max(\frac{1}{1-\alpha}, \frac{1}{\alpha})$$

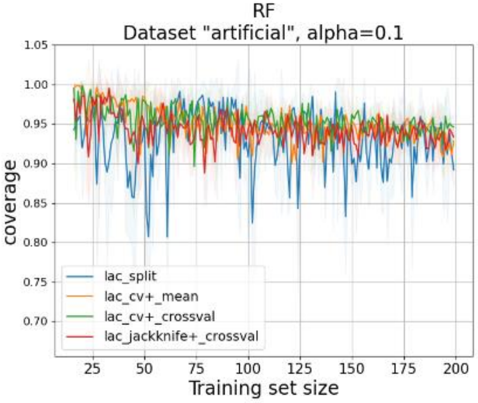
  - necessary for determine the 1-alpha quantile

- smallest training set size $|\mathcal{Y}|$

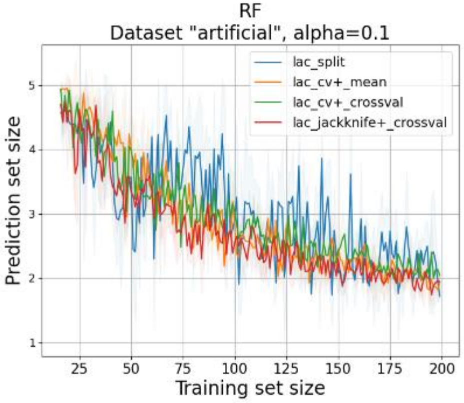
- for 5 classes and alpha=0.1

$$n \geq \frac{1}{1-\alpha} + 1 + |\mathcal{Y}| = 16$$

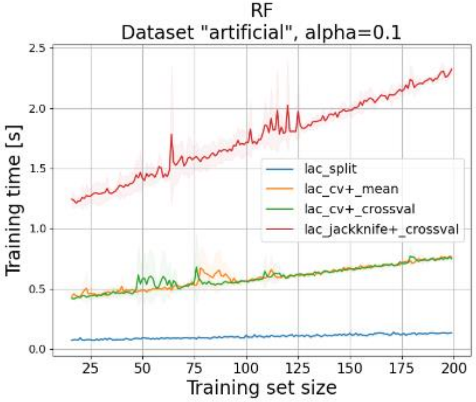
Additional Experiments – very small datasets



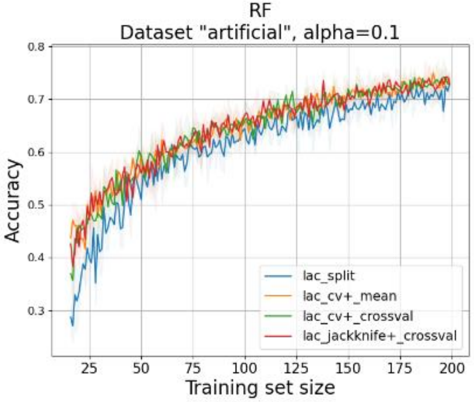
(a) Coverage



(b) Average prediction set size

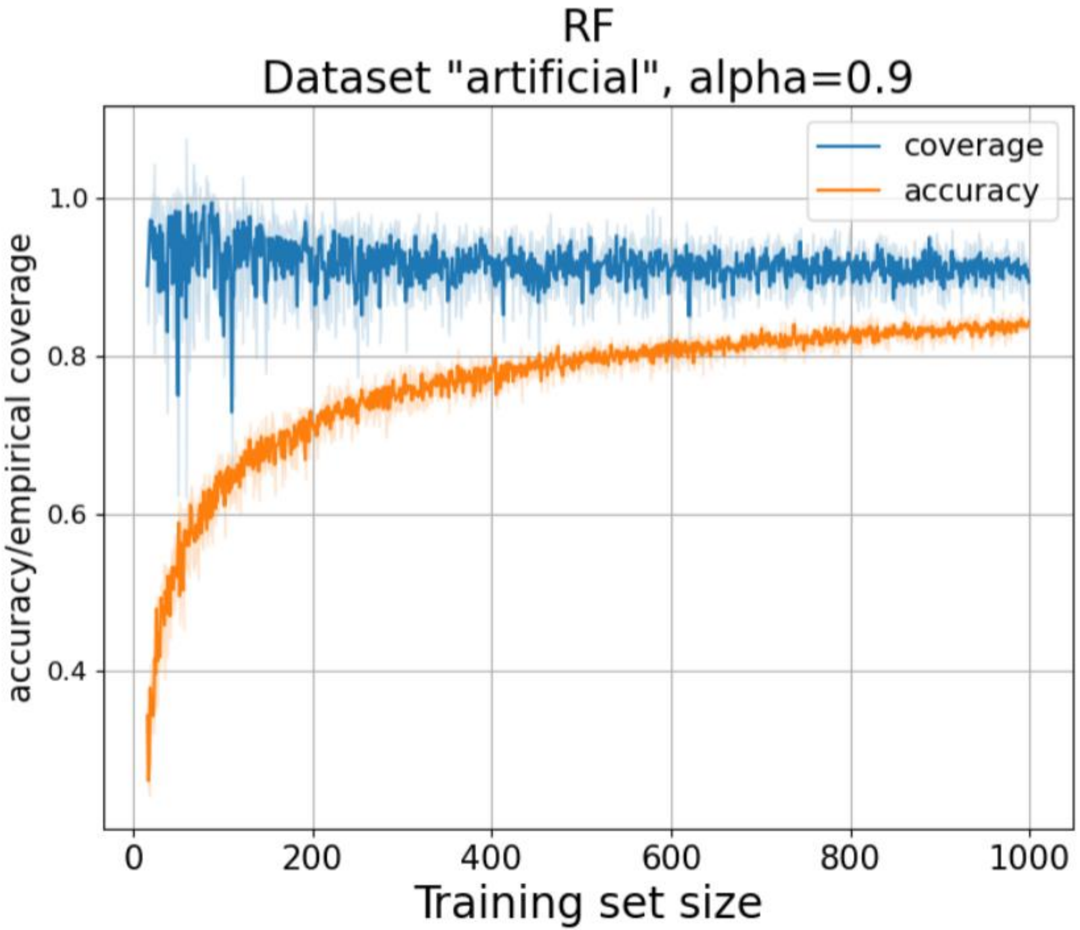


(c) Training time



(d) Accuracy

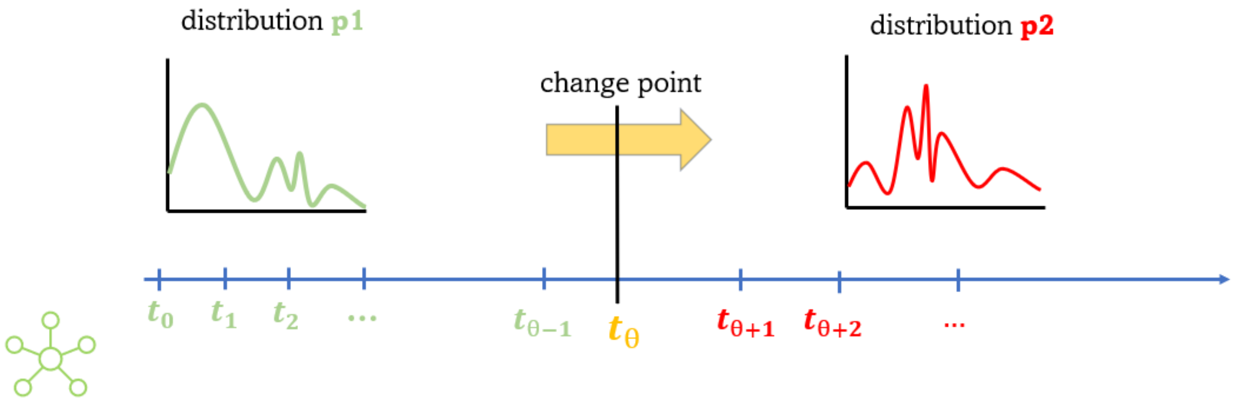
Additional Experiments – direct comparison acc and cov



Conformal Change Point Detection

- Based on Vovks Paper:
  - Retrain or not retrain: conformal test martingales for change-point detection [8] (2021)

Scenario:



t_0 : train model on **p1**

$t_1 - t_{...}$: online evaluation

- calculate CTM M_i at each timepoint

Changepoint detected if:

- M_i > threshold

Conformal Test Martingales CTM

- data sequence: $$(z_1, z_2, \ldots, z_{n+1})$$
- conformity measure: $$s_i = S(z_i)$$
- p-values: $$p_{n+1} = \frac{|\{i|s_i > s_{n+1}\}| + \theta_{n+1}|\{i|s_i = s_{n+1}\}|}{n}$$
- CTM: $$M_{n+1} = F(p_1, p_2, \ldots, p_{n+1})$$

F is a betting martingale function

Betting Martingale

- for all sequences: $p_1, p_2, ..., p_n \ , \ \in [0,1]^n$

$$\int_0^1 F(p_1, ..., p_n, p_{n+1}) du = F(p_1, ..., p_n)$$

- therefore, the following property holds:

$$\mathbb{E}(M_n | S_1, ..., S_{n+1}) = S_{n-1}$$

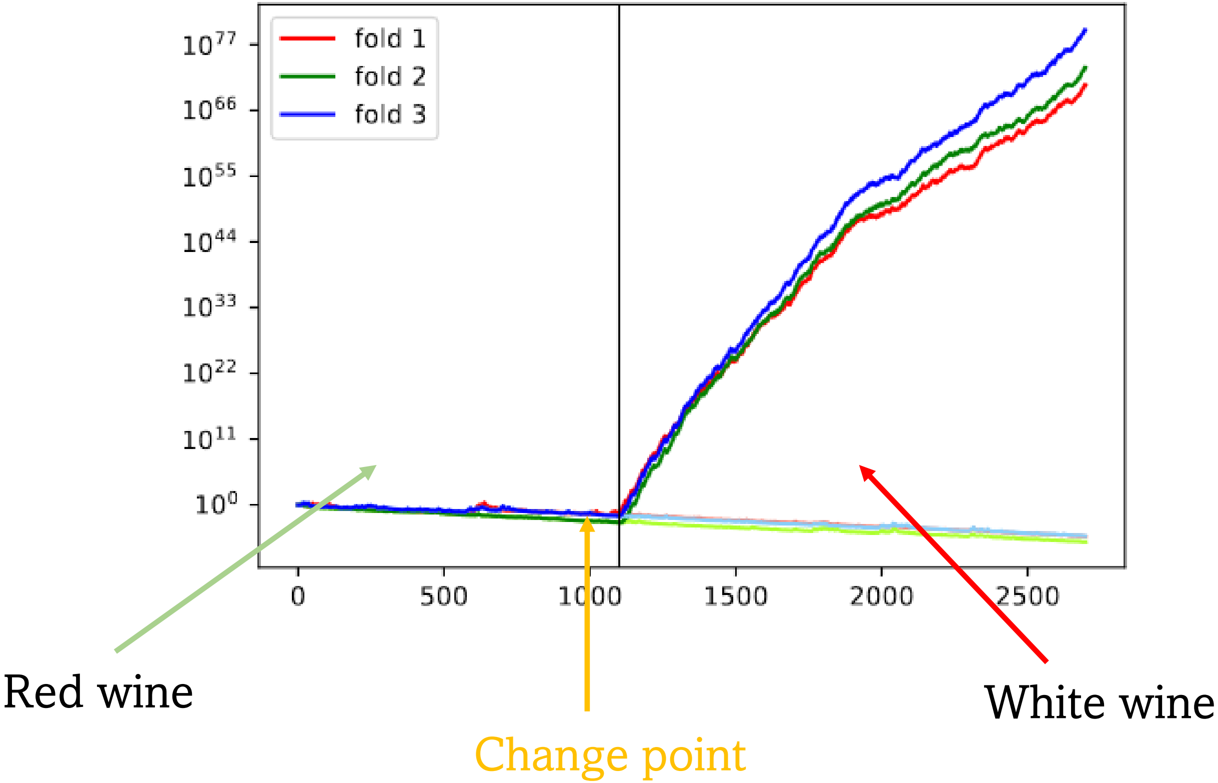
CTM: betting martingale function – Simple Jumper

$$F(p_1, p_2, ..., p_{n+1}) = \int \prod_{i=1}^{n} f_{\epsilon_i}(p_i)\mu(d(\epsilon_1, \epsilon_2, ...))$$

with

$$f_{\epsilon_i}(p) = 1 + \epsilon(p - 0.5)$$

CTM: Vovks results



CTM - properties

- under iid. p-values are uniformly distributed in [0,1]

- Simple Jumper martingales fulfil Ville's inequality:

$$P(\exists i : M_i \geq c) \leq \frac{1}{c}$$

- M_i are equal under iid.
- CTM are equal under iid.
    - **M_i > c with false alarm rate of alpha = 1/c if iid. violated**

CTM: Vovks results

- CTM controls the false alarm rate
  - controlling c

- not as efficient as other methods (CUMSUM, Shiryaev-Roberts procedure)

- Empirically detect change point for clear changepoint and clear to separate distributions after 20-30 datapoints

- best conformity measure function: $$s_i = y_i - \hat{y}_i$$

  - Even better then L1-Norm
  - Possible because real label is available in online setting

PERMAD Dataset

- 41 oncological patients
- total 647 measurements
- different number of measurements per patient (min: 4, max: 46)
- time distributed irregularly and different for each patient
- 91 Features

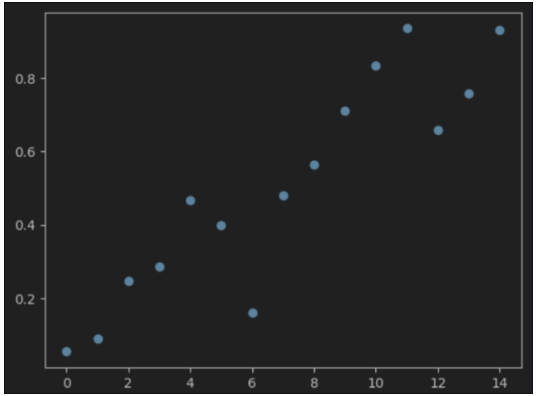
- CT scans at irregular intervals
- progression / non-progression

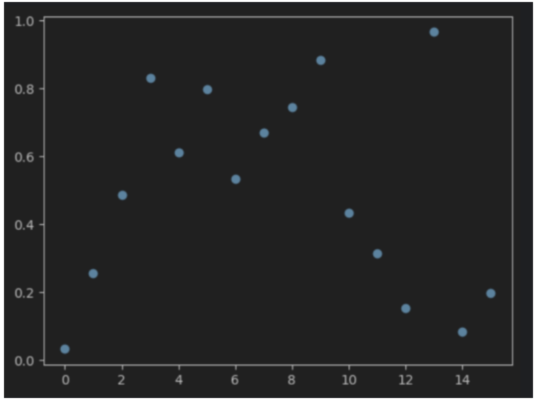
- Task: predict the change-point, from non-progression to progression

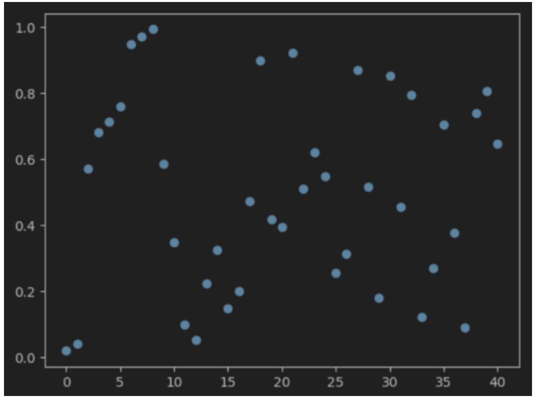
PERMAD Dataset - Idea

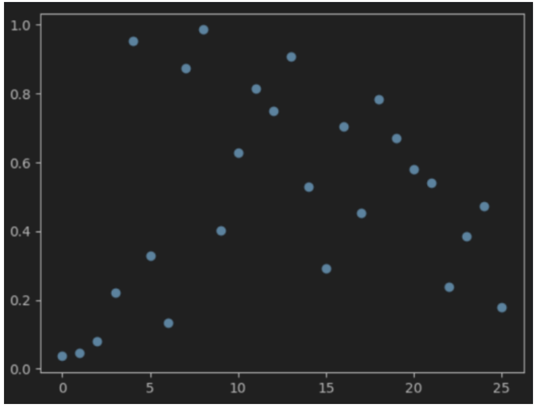
- No ground truth
  - Learn an unsupervised representation of "non-progression" and "progression"
- What is p1 , what p2
  - Solution: fist datapoint of each patient is "non-progression" , last progression
- Model: Autoencoder
- Training: Two AE one in the first one on the last datapoints of all patients
- Non-conformity measure
  - difference between t_0 and t_i (reconstructions / embeddings)
  - use both AEs alone and in combination
  - diff, L1, L2, cross-entropy
- Calculate CTM as search for a method that shows a changepoint

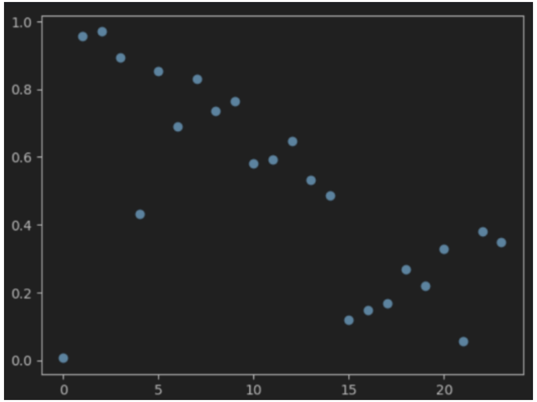
PERMAD Dataset – Results (L1 – embeddings)

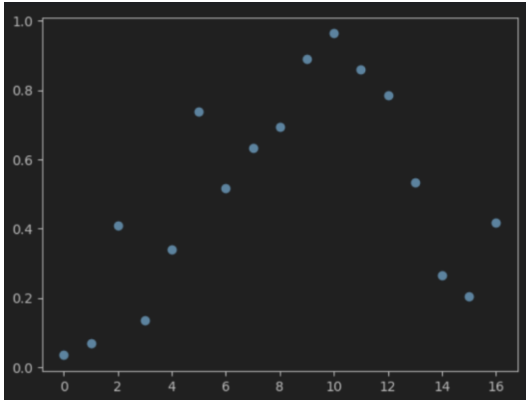




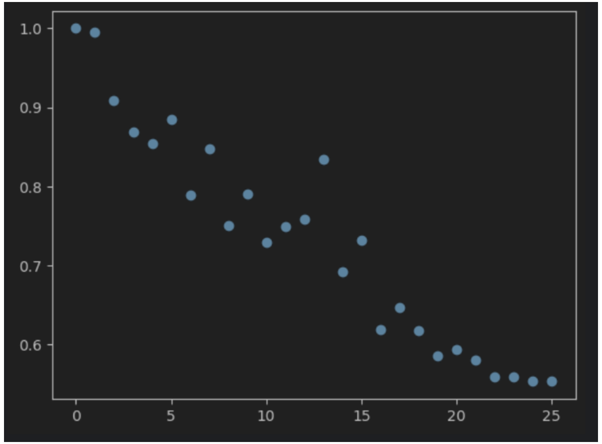


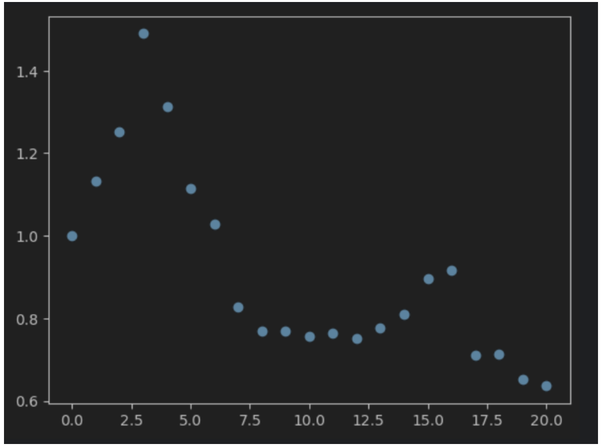


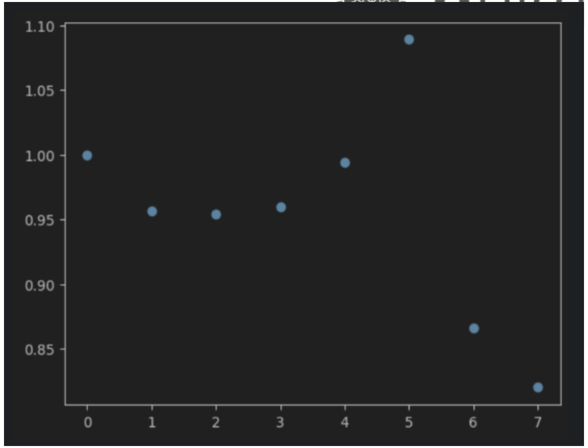


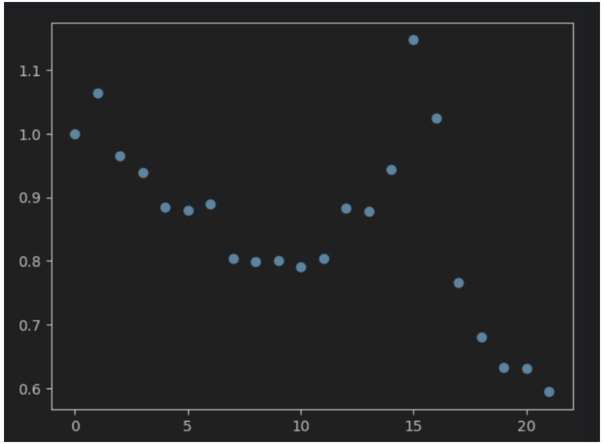


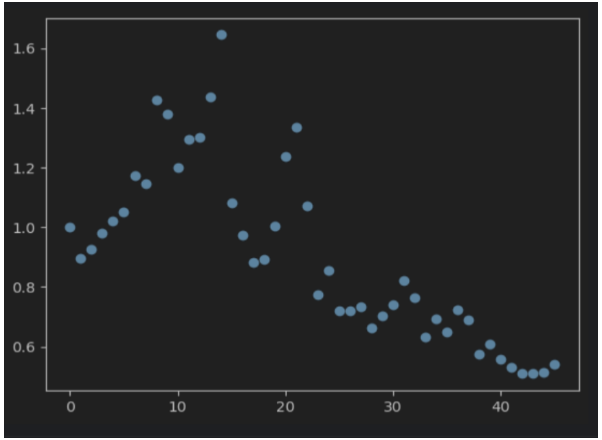
PERMAD Dataset – Results (L2 – embeddings)

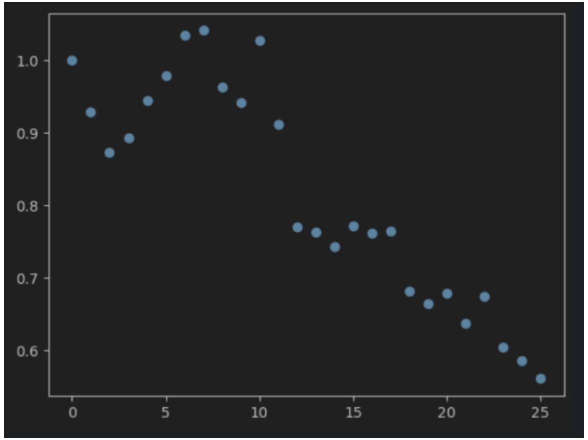




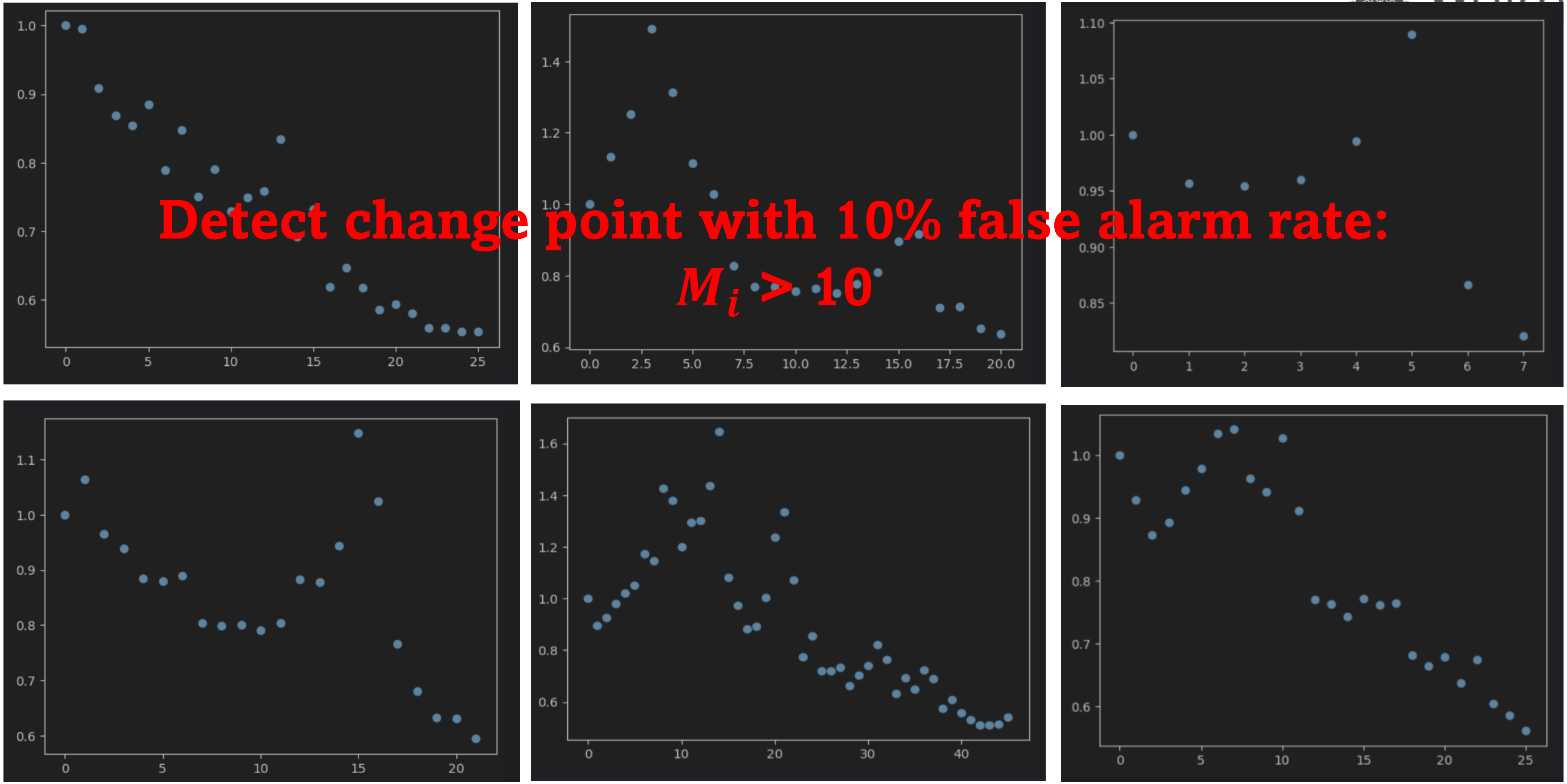








PERMAD Dataset - Results



Detect change point with 10% false alarm rate:

$$M_i > 10$$

PERMAD Dataset - Problems

- CTM and the given task does not match
- No ground truth / no supervised task
- CTM is not efficient (requires in bast case scenarios >20 datapoints of the other distribution)

- Questionable if "non-progression" / "progression" distribution exists ?
- Are intersubject differences bigger then "progression"/ "non-progression"?
- Maybe more than one change point?
  - available features reflect more than just the oncological status
  - Subject: can become sick, co-medication, …
  - smooth transition, not an abrupt change

Quellen

[1] Vovk, V. (2012). Conditional validity of inductive conformal predictors. In Asian conference on machine learning, Vol. -, pp. 475–490. PMLR.

[2] Angelopoulos, A. N., & Bates, S. (2021). A gentle introduction to conformal prediction and distribution-free uncertainty quantification. arXiv preprint arXiv:2107.07511

[3] Romano, Y., Sesia, M., & Candes, E. (2020). Classification with valid and adaptive coverage. Advances in Neural Information Processing Systems, 33, 3581–3591

[4] Lei, J., & Wasserman, L. (2014). Distribution-free prediction bands for non-parametric regression. Journal of the Royal Statistical Society Series B: Statistical Methodology, 76 (1), 71–96.

[5] Lei, J. (2014). Classification with confidence. Biometrika, 101 (4), 755–769

[6] Koklu, M., & Ozkan, I. A. (2020). Multiclass classification of dry beans using computer vision and machine learning techniques. Computers and Electronics in Agriculture, 174, 105507.

[7] Fontana, Matteo, Gianluca Zeni, and Simone Vantini. "Conformal prediction: a unified review of theory and new challenges." Bernoulli 29.1 (2023): 1-23.

[8] Vovk, V., Petej, I., Nouretdinov, I., Ahlberg, E., Carlsson, L., & Gammerman, A. (2021). Retrain or not retrain: Conformal test martingales for change-point detection. In Conformal and Probabilistic Prediction and Applications, pp. 191–210. PMLR.