Sequential Conformal Prediction for Time Series

Chen Xu, Yao Xie

H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology

ICML DFUQ Workshop, July 2022

Outline

Prediction intervals for time series

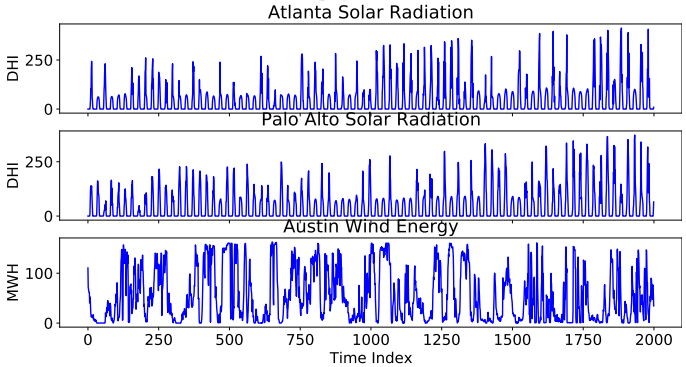
- Problem setup
- Sequential conformal prediction
  - Vanilla sequential conformal (EnbPI)
  - Sequential Predictive Conformal Inference (SPCI)
- Extension: Conformal prediction set

Conformal prediction for time-series. Xu and **X**.
`https://arxiv.org/abs/2010.09107`

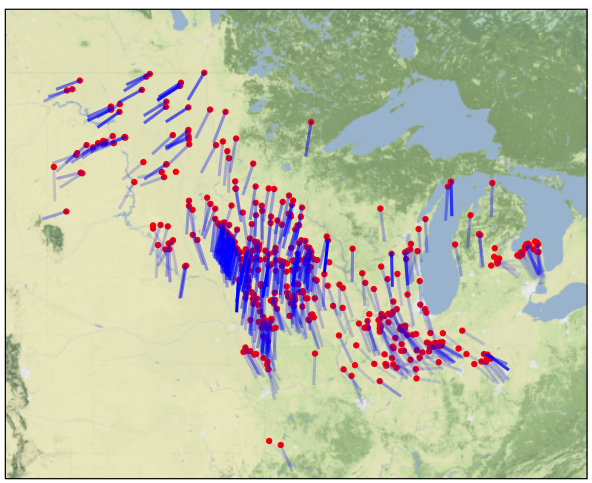
ICML 2021 (Long presentation).

Time series data

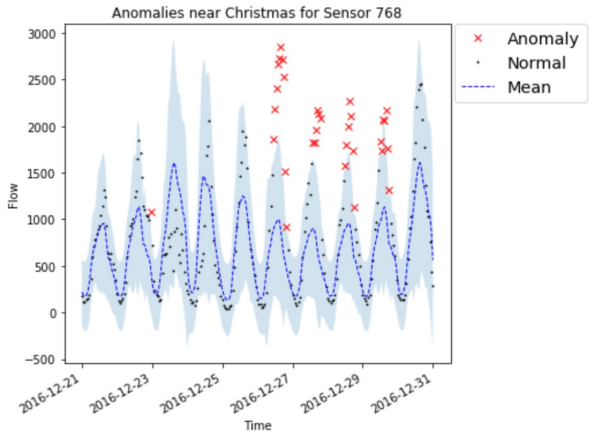
Solar energy prediction



Wind power prediction



Real-time anomaly detection



ICU sequential data prediction

Problem setup

- *Conformal prediction*: constructing prediction intervals that attain valid coverage in finite samples, without making distributional assumptions.
- Time series conformal prediction:

$$Y_t = f_t(X_t) + \epsilon_t, \quad t = 1, 2, \ldots$$

$$Y_t \in \mathbb{R}, \quad X_t \in \mathbb{R}^d, \quad \epsilon_t \sim F$$

- Features X_t can be either exogenous time-series and/or the history of Y_t, e.g.,

$$X_t = (Y_{t-1}, \ldots Y_{t-p}, Z_t)$$

- Given a prediction algorithm $\hat{f}_t$ trained using data $\{X_t, Y_t\}, t = 1, \ldots, T$, generate prediction for X_t, $t > T$

Goal

- Goal: Quantify the uncertainty of time series prediction algorithm $\hat{f}_t(X_t)$, $t > T$

- Construct prediction intervals $\widehat{C}_t^{\alpha}, t > T$, with pre-specified significance level $\alpha > 0$
  - *conditional* coverage guarantee

$$P(Y_t \in \widehat{C}_t^{\alpha} | X_t) \geq 1 - \alpha.$$

  - *marginal* coverage guarantee:

$$P(Y_t \in \widehat{C}_t^{\alpha}) \geq 1 - \alpha.$$

Non-sequential conformal inference

Requires data exchangeability

- Split conformal (Vovk et al. 2005)

Training	holdout	X_{n+1}

$n/2$ training $(X_i, Y_i) \rightarrow \hat{f}(\cdot)$

Residuals on $n/2$ holdout: $R_i = \left|Y_i - \hat{f}(X_i)\right|$

$\hat{C}_n(X_{n+1}) = \hat{f}(X_{n+1}) \pm$ quantile of $\{R_i\}_{i=1}^n$

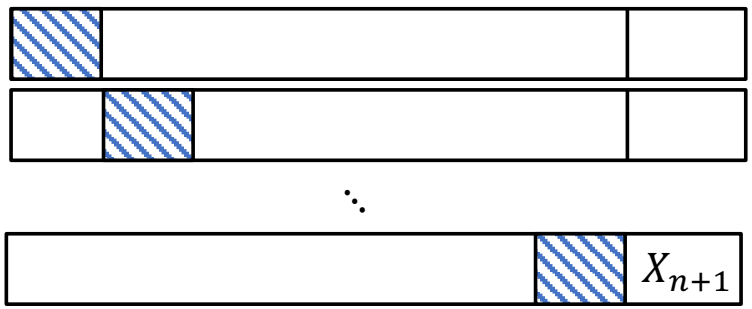
- Full conformal – avoid splitting
(Vovk et al. 2005), Lasso (Lei 2019)

Training	(X_{n+1}, y)

$(X_1, Y_1) \ldots (X_n, Y_n), (X_{n+1}, y) \rightarrow \widehat{f_y}(\cdot)$

- Jackknife+ (Barber et al. 2021)

Avoid splitting by consider leave-one-out



$\hat{f}_{-i}$ fitted leaving out (X_i, Y_i)

Using empirical distribution LOO residuals

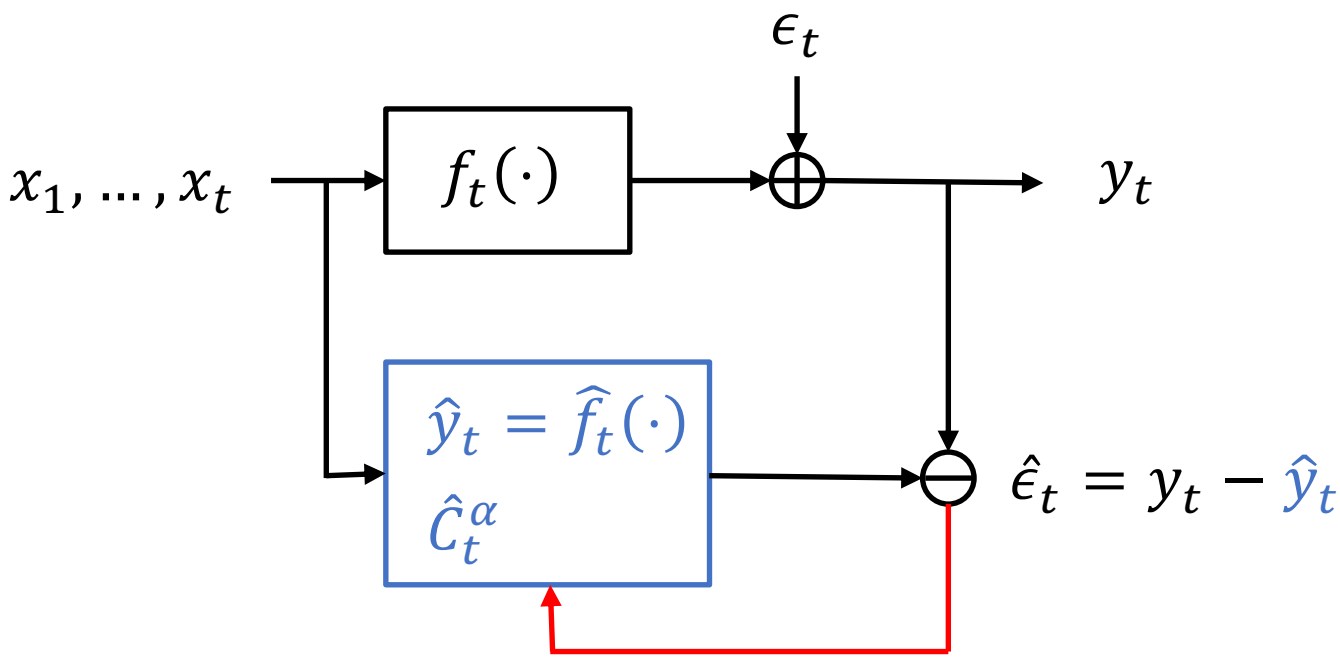
- Conformalized quantile regression
(Romano et al. 2019)
  - Based on empirical distribution of residuals
  - Conditional quantile regression (others are conditional mean regression)
  - Handle heteroscedasticcity

Beyond exchangeability

(Potentially) applicable to sequential data

- (Tibshirani et. al, 2019) *Weighted exchangeability*
  - Handle covariance shift
  - Requires full knowledge of change in distribution
- (Barber et al., 2022)
  - Weights are fixed (rather than data-dependent)
  - for unknown violation of exchangeability
- (Podkopaev, Ramdas 2021)
  - reweighting can also deal with label shift
- (Lei, Candes 2021)
  - reweighting for causal inference
- (Gibbs, Candes 2021)
  - Adjust α_t, by comparing empirical coverage with target level $(1-\alpha)$, using stochastic gradient descent

Sequential conformal inference



- Data not exchangeable
- Feedback available:
  Algorithm predicts $\hat{Y}_t \rightarrow$ True Y_t reveals $\rightarrow$ Feedback $\hat{\epsilon}_t$
- Nature can generate temporally correlated ϵ_t with unknown pdf

Sequential conformal inference

- Vanilla version: `EnbPI`
  Based on empirical distribution of residuals
  - Based on tail of $\{\hat{\epsilon}_i\}, i = 1, \ldots, t-1$
  - Guarantee for i.i.d., weak dependence, α-mixing

- Sequential Predictive Conformal Inference (SPCI):
  Exploiting temporal dependence of residuals
  - **Quantile regression** to get $\mathbb{P}\{\hat{\epsilon}_t > x | \hat{\epsilon}_{t-1}, \ldots, \hat{\epsilon}_{t-w}\}$
  - Guarantee for stationary time-series – allowing strong dependence

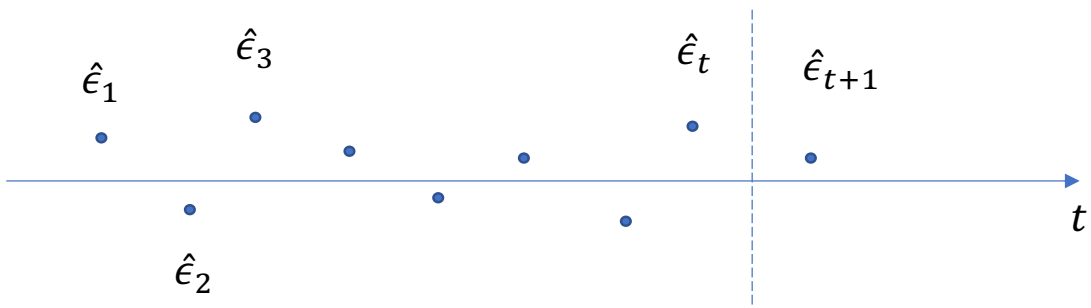
- Handling non-stationarity and heteroskedasticity
  Treating time (and periodicity) as predictor

Sequential conformal inference: Vanilla version

- $\hat{f}_t$: Prediction algorithm trained using data $\{X_i, Y_i\}, i = 1, \ldots, T$ for predicting at X_t, $t > T$
- Prediction residual

$$\hat{\epsilon}_t = Y_t - \hat{f}_t(X_t)$$

- Set of past prediction residuals $\mathcal{E}_{t-1} := \{\hat{\epsilon}_i\}_{i=t-1,\ldots,t-w}$



Conformal prediction for time-series. Xu, **X**. ICML 2021. (Long Talk.)

EnbPI

(Xu and **X**., 2021)

Prediction interval at level $(1-\alpha)$

$$\widehat{C}_t^{\alpha} = [\hat{f}_t(X_t) + Q_{\beta^*}(\mathcal{E}_{t-1}), \hat{f}_t(X_t) + Q_{1-\alpha+\beta^*}(\mathcal{E}_{t-1})],$$
$$\beta^* := \underset{\beta\in[0,\alpha]}{\arg\min} \ \left(Q_{1-\alpha+\beta}(\mathcal{E}_{t-1}) - Q_{\beta}(\mathcal{E}_{t-1})\right).$$

Q_α computes empirical α quantile of $\mathcal{E}_{t-1} := \{\hat{\epsilon}_i\}_{i=t-1,\ldots,t-w}$

- Prediction intervals enjoy marginal coverage asymptotically
- Theoretical guarantees hold without exchangability assumption

Theoretical guarantee: EnbPI

Holds for $f_t(X_t) = f(X_t)$:

- WLOG, analyze $t = T + 1$; can extend to $t > T + 1$
- Assumption 1 (Data regularity): Error process $\epsilon_1, \epsilon_2, \ldots$
  - stationary and strongly mixing
  - sum of mixing coefficients bounded by M
  - true CDF F is Lipschitz with constant $L > 0$
- Assumption 2 (Estimation quality)
  There exists a real sequence $\{\delta_T\}_{T \geq 1}$ converging to zero such that

$$\sum_{t=1}^{T} (\hat{f}_t(X_t) - f(X_t))^2 / T \leq \delta_T^2,$$

Theoretical guarantee (cont.)

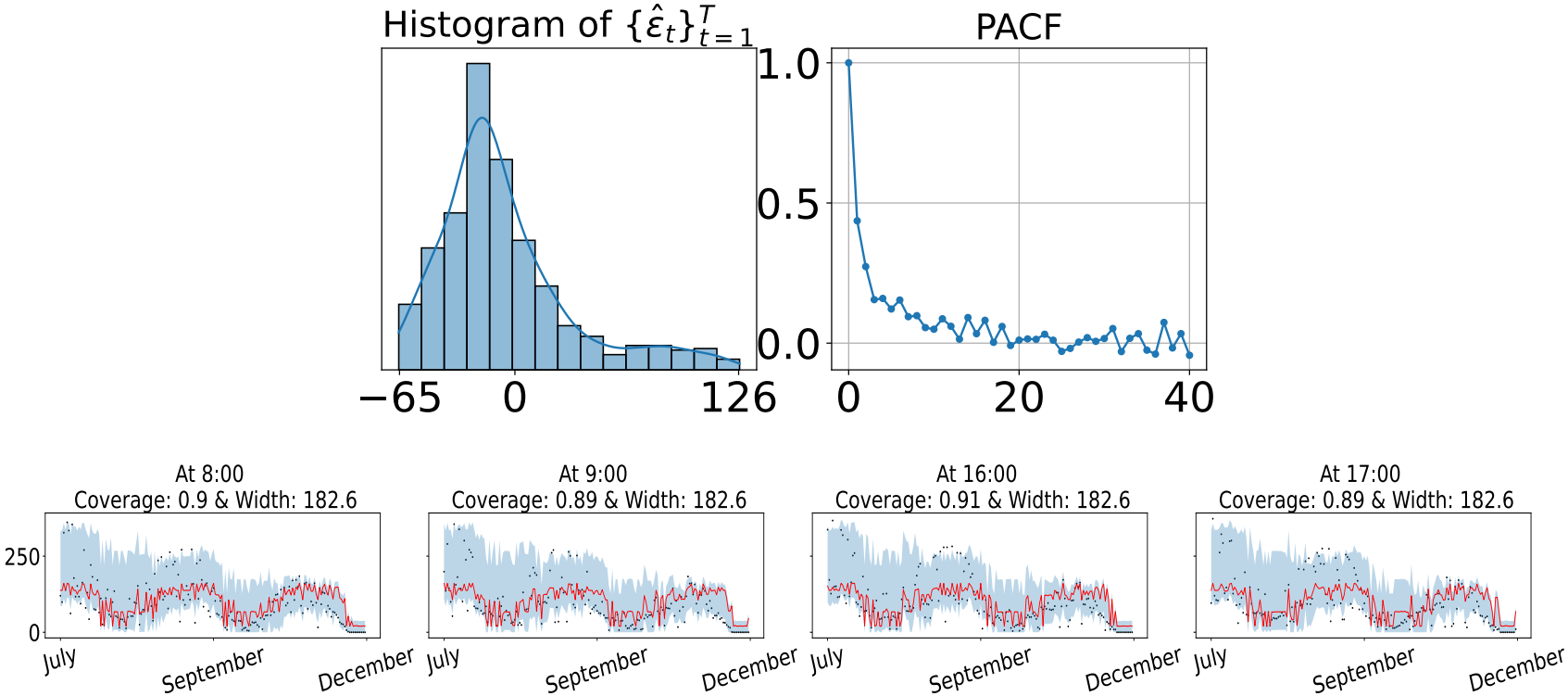
- Given a training size T and $\alpha \in (0, 1)$,

$$|\mathbb{P}(Y_{T+1} \notin \widehat{C}^{\alpha}_{T+1}) - \alpha| \leq C((\log T/T)^{1/3} + \delta_T^{2/3})$$

Implications

- Factor $(\log T/T)^{1/3}$ comes from assuming α-mixing errors, different error assumptions (e.g., independent, stationary, etc.) yield different rates

What do residuals look like? Solar power

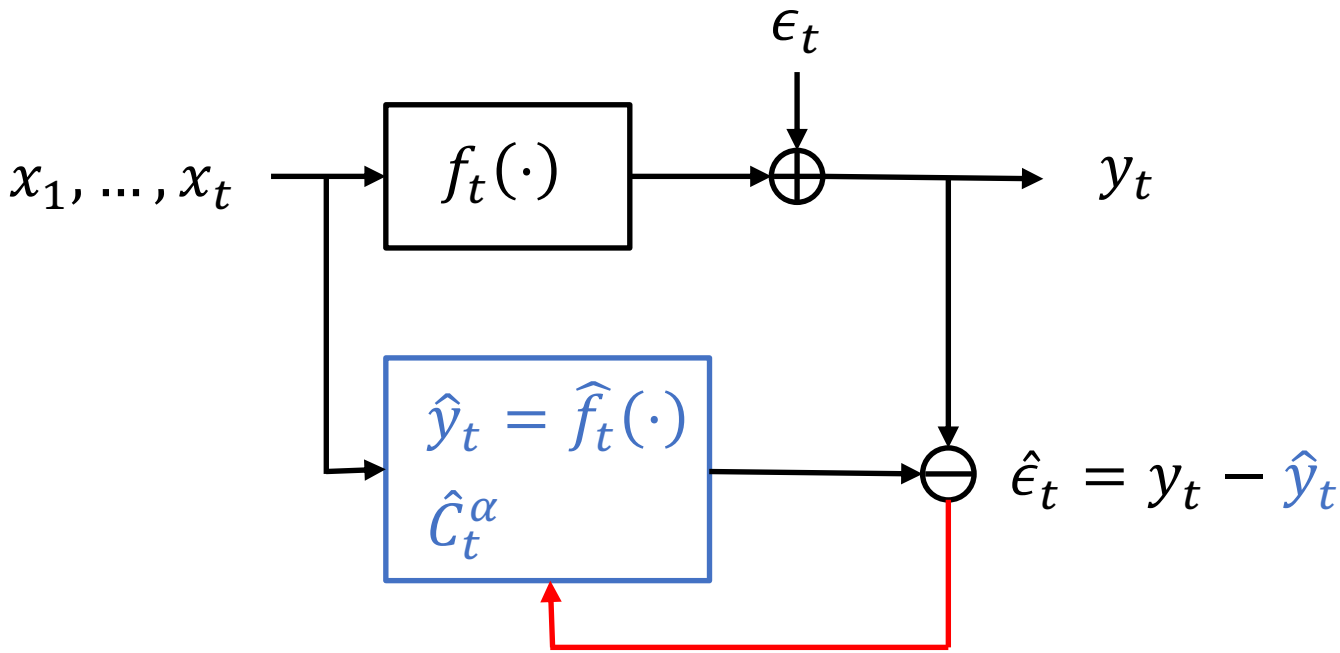


Bottom figure: `EnbPI` coverage conditioning at each hour.

- Asymmetric residual distribution
- Residuals have temporal correlation

Revisit: Sequential conformal inference

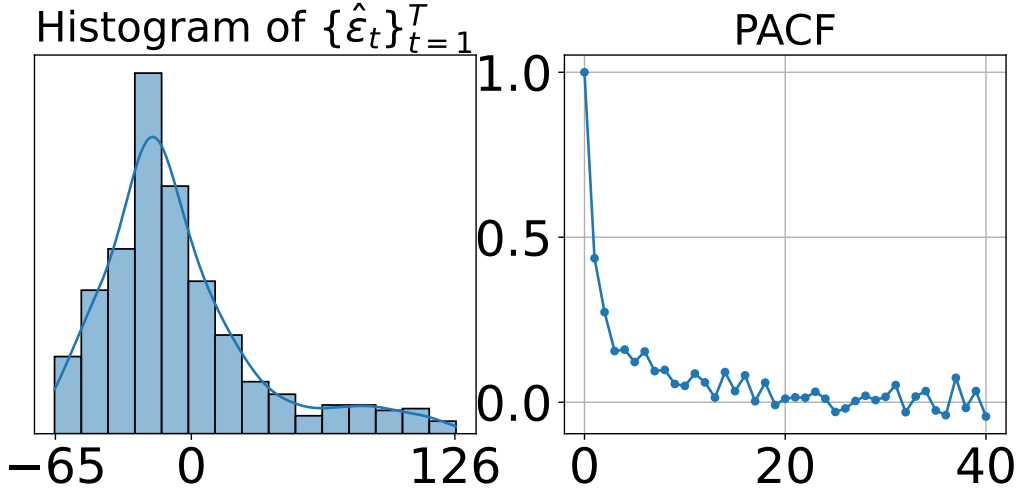
Can we do better using properties of $\{\hat{\epsilon}_t\}$?



- Feedback available:
  Algorithm predicts $\hat{Y}_t \rightarrow$ True Y_t reveals $\rightarrow$ Feedback $\hat{\epsilon}_t$

What's in prediction residuals?

$$\hat{\epsilon}_t = Y_t - \hat{Y}_t = \underbrace{f_t(X_t) - \hat{f}_t(X_t)}_{\text{prediction error}} + \underbrace{\epsilon_t}_{\text{"nature"}}$$



$\hat{\epsilon}_t$ may be temporally correlated:

- Prediction error, e.g., model is biased
- "nature" generates correlated noise ϵ_t

Time series generally not exchangeable

- What happens without exchangeability?

simplicity, we assume that there are almost surely no ties among the scores $V_1, \ldots, V_{n+1}$, so that we can work with sets rather than multisets (our arguments apply to the general case as well, but the notation is more cumbersome).

Denote by E_v the event that $\{V_1, \ldots, V_{n+1}\} = \{v_1, \ldots, v_{n+1}\}$, and consider

$$\mathbb{P}\{V_{n+1} = v_i \,|\, E_v\}, \; i = 1, \ldots, n+1. \tag{13}$$

Denote by f the probability density function[4] of the joint distribution $V_1, \ldots, V_{n+1}$. Exchangeability implies that

$$f(v_1, \ldots, v_{n+1}) = f(v_{\sigma(1)}, \ldots, v_{\sigma(n+1)})$$

for any permutation σ of the numbers $1, \ldots, n+1$. Thus, for each i, we have

$$\begin{aligned}\mathbb{P}\{V_{n+1} = v_i \,|\, E_v\} &= \frac{\sum_{\sigma:\sigma(n+1)=i} f(v_{\sigma(1)}, \ldots, v_{\sigma(n+1)})}{\sum_{\sigma} f(v_{\sigma(1)}, \ldots, v_{\sigma(n+1)})} \\ &= \frac{\sum_{\sigma:\sigma(n+1)=i} f(v_1, \ldots, v_{n+1})}{\sum_{\sigma} f(v_1, \ldots, v_{n+1})} \\ &= \frac{n!}{(n+1)!} = \frac{1}{n+1}.\end{aligned} \tag{14}$$

Without exchangeabilty, this does not hold. But assuming stationary, we can fit a "fix" predictive model for this.

This shows that the distribution of $V_{n+1}|E_v$ is uniform on the set $\{v_1, \ldots, v_{n+1}\}$, i.e.,

$$V_{n+1}|E_v \sim \frac{1}{n+1} \sum_{i=1}^{n+1} \delta_{v_i},$$

and it follows immediately that

$$\mathbb{P}\left\{V_{n+1} \leq \text{Quantile}\left(\beta; \frac{1}{n+1} \sum_{i=1}^{n+1} \delta_{v_i}\right) \,\middle|\, E_v\right\} \geq \beta.$$

"Conformal prediction under covariate shift," Tibshirani, Barber, Candès, Ramdas, 2020

Exploit temporal dependence in prediction residuals?

- Dependence of residuals means that $\{\hat{\epsilon}_{t-1}, \ldots, \hat{\epsilon}_1\}$ contain information about $\hat{\epsilon}_t$

$$\hat{\epsilon}_t | \{\hat{\epsilon}_{t-1}, \ldots, \hat{\epsilon}_{t-w}\} \overset{\mathsf{d}}{\neq} \hat{\epsilon}_t$$

- What's one typical characteristic to time-series?

(Stationarity): $(\hat{\epsilon}_{t-w}, \ldots, \hat{\epsilon}_t) \overset{\mathsf{d}}{=} (\hat{\epsilon}_{t-w+d}, \ldots, \hat{\epsilon}_{t+d}), \forall w, d$

- We can build a predictive model for conditional tail probability using quantile regression

$$\mathbb{P}\{\hat{\epsilon}_t > x | \hat{\epsilon}_{t-1}, \ldots, \hat{\epsilon}_{t-w}\}, \quad \text{for given } x$$

Proposed algorithm: SPCI

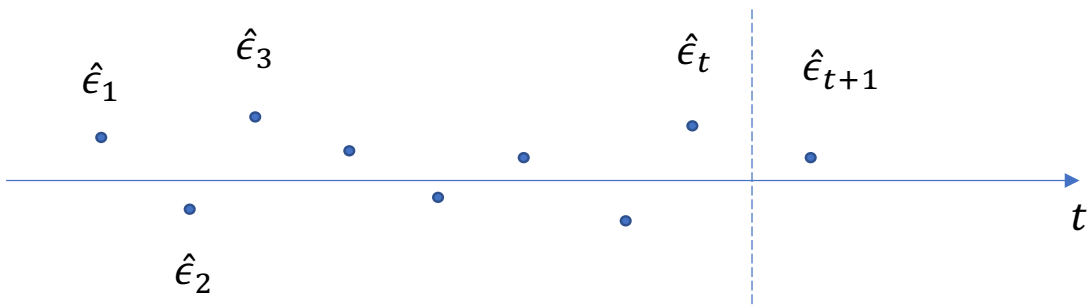
`Sequential Predictive Conformal Inference (SPCI)`

- Idea: Achieve adaptivity by predicting residual quantile from past observed residuals sequentially:

$$\mathbb{P}\{\hat{\epsilon}_t > x | \hat{\epsilon}_{t-1}, \ldots, \hat{\epsilon}_{t-w}\}$$

  Use predicted quantile to estimate $\widehat{C}_t^{\alpha}$

- Can use *quantile regression*, e.g., random forest based to estimate above

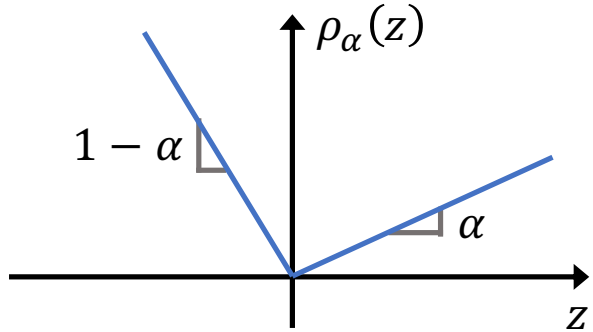


Quantile regression

- Quantile regression estimates conditional quantile functions from data

$$\hat{Q}_\alpha(x) = f(x, \hat{\theta}), \quad \hat{\theta} = \arg\min_\theta \frac{1}{n} \sum_{i=1}^{n} \rho_\alpha(y_i, f(x_i, \theta)) + R(\theta)$$

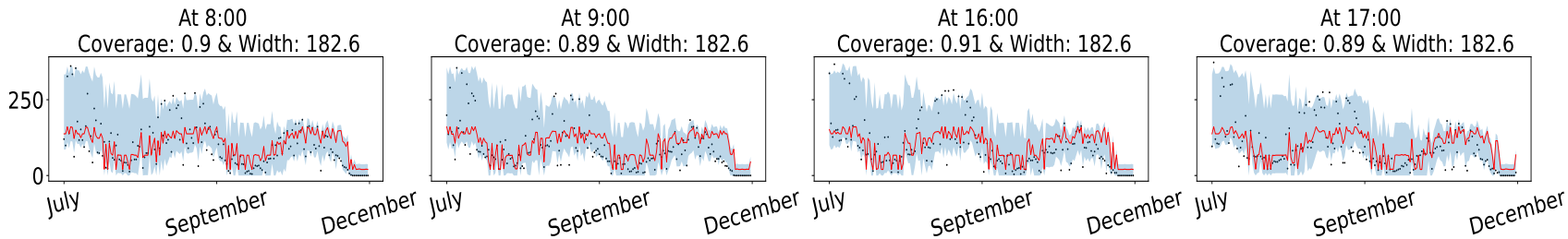
$f(x, \theta)$: Quantile regression function
ρ_α: Check function or pinball loss
$R(\theta)$: A potential regularizer



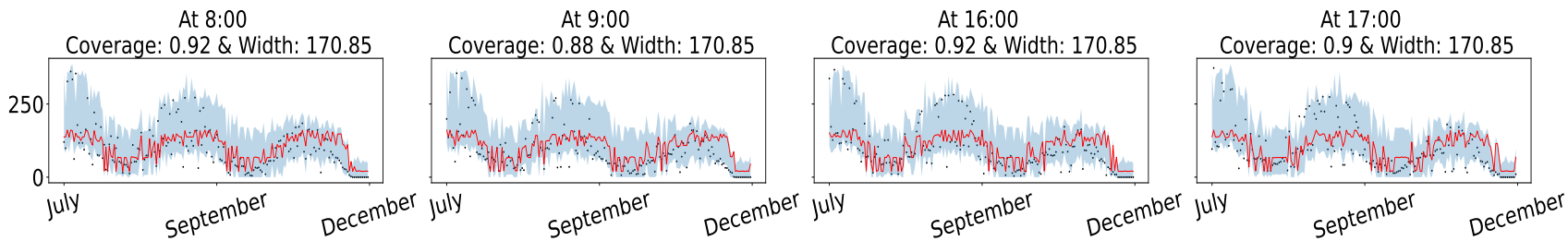
(Keonker, Bassettm 1978) (Romano, Patterson, Candés, 2019)

Back to solar power

- Coverage: SPCI $\approx$ EnbPI
- Interval width: SPCI $<$ EnbPI



(a) EnbPI conditional coverage and width at each hour



(b) SPCI conditional coverage and width at each hour

Handling non-stationarity and heteroskedasticity

Non-stationarity

- True model $Y_t = f_t(X_t) + \epsilon_t$, true model depends on t
- If we define $X'_t := [g(t), X_t]$
- Then $\hat{f}_t(X'_t)$ can handle a variety of non-stationary models:
  $g(t) = ct$: trend
  $g(t) = \text{mod}(t, 12)$: periodicity with period 12

Heteroskedasticity

- $Y_t = f(X_t) + \sigma(X_t)\epsilon_t$
- Jointly estimate both $f_t(X_t)$ and $\sigma(X_t)$ using normalized residuals

$$\hat{\epsilon}_t := (Y_t - \hat{f}_t(X_t))/\hat{\sigma}(X_t).$$

Simulated example:

Compare EnbPI and SPCI on stationary data

- State-space model:

$$Y_t = \alpha Y_{t-1} + \epsilon_t$$

- AR(1) error: $\epsilon_t = \beta \epsilon_{t-1} + v_t, v_t \sim N(0, 0.1)$
- Fix $\alpha = \beta = 0.9$.
- Quantile regression with random forest

Training fraction	50%	60%	70%	80%
EnbPI coverage	0.90	0.86	0.87	0.88
SPCI coverage	0.92	0.97	0.94	0.93
EnbPI width	16.31	15.86	16.54	16.17
SPCI width	5.71	4.97	4.73	5.44

Simulated example:

Non-stationary time-series

$$Y_t = f_t(X_t) + \epsilon_t, \text{ same AR(1) } \epsilon_t.$$

- Non-stationary model

$$f_t(X_t) = g(t)h(X_t),$$
$$g(t) = t' \sin(2\pi t'/12), \quad t' = \text{mod}(t, 12)$$
$$h(X_t) = (|\beta^T X_t| + (\beta^T X_t)^2 + |\beta^T X_t|^3)^{1/4}$$
$$X_t \in \mathbb{R}^{100}, X_t = [Y_{t-1}, \ldots, Y_{t-100}].$$

- Target coverage 0.9

Training fraction	50%	60%	70%	80%
EnbPI coverage	0.90	0.90	0.91	0.90
SPCI coverage	0.91	0.92	0.92	0.94
EnbPI width	25.13	25.43	25.56	25.16
SPCI width	11.35	11.17	11.23	11.31

Simulated example:

Heteroskedastic time-series

$$Y_t = f_t(X_t) + \epsilon_t, \text{ same AR(1) } \epsilon_t.$$

- Heteroskedastic model
  $f(X_t) = h(X_t),\ \sigma(X_t) = \mathbf{1}^T X_t.$
  $X_t \in \mathbb{R}^{20}$, with i.i.d. entries from Uniform$[0, e^{0.2\text{mod}(t,200)})$.
- Target coverage 0.9
- Only use `EnbPI` because it is faster
- `SPCI` can reduce interval widths

Training fraction	10%	20%	30%	40%	50%	60%	70%	80%
Coverage	0.88	0.89	0.89	0.90	0.90	0.90	0.90	0.90
Width	5.00	5.16	5.08	4.79	4.91	4.80	4.99	5.05

Solar data:

Compare EnbPI and SPCI

Table: EnbPI and SPCI — "non-stationary" versions add hourly variables to feature X_t to handle time non-stationarity. Target coverage is 0.9.

Training fraction	50%	60%	70%	80%
EnbPI coverage	0.88	0.88	0.87	0.86
EnbPI non-stationary coverage	0.89	0.88	0.88	0.90
SPCI non-stationary coverage	0.89	0.90	0.89	0.90
EnbPI width	110.01	103.32	106.56	100.31
EnbPI non-stationary width	73.20	54.18	52.90	50.37
SPCI non-stationary width	58.83	48.55	42.24	44.61

Solar data:

Compare with adaptive CI

- AdaptiveCI (Gibbs, Candés 2022) vs. `EnbPI` and `SPCI`
- Under non-stationary X_t, `EnbPI` and `SPCI` significantly improve.

Training fraction	50%	60%	70%	80%
AdaptiveCI coverage	0.90	0.89	0.88	0.91
`EnbPI` coverage	0.88	0.88	0.87	0.86
`SPCI` coverage	0.89	0.90	0.89	0.90
AdaptiveCI width	72.65	68.20	70.39	66.21
`EnbPI` width	110.01	103.32	106.56	100.31
`SPCI` width	70.95	77.33	67.19	70.66

(a) Original $X_t, \alpha = 0.1$.

Training fraction	50%	60%	70%	80%
AdaptiveCI coverage	0.90	0.91	0.90	0.90
`EnbPI` coverage	0.89	0.88	0.88	0.90
`SPCI` coverage	0.89	0.90	0.89	0.90
AdaptiveCI width	62.56	60.61	57.87	59.62
`EnbPI` width	73.20	54.18	52.90	50.37
`SPCI` width	58.83	48.55	42.24	44.61

(b) Considering non-stationarity, $\alpha = 0.1$.

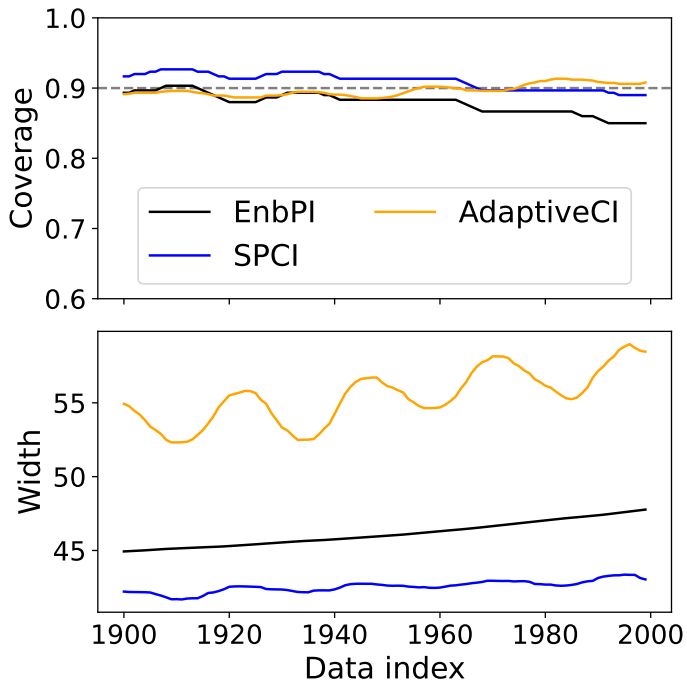


Figure: Rolling coverage considering non-stationarity; Training Fraction = 80% or 1600 obs.

Solar Data

Compare with weighted non-exchangeable conformal

- Nex-CP WLS (Barber et al. 2022) vs. `EnbPI` and `SPCI`
- Both `EnbPI` and `SPCI` yield narrower intervals regardless of X_t.

Training fraction	50%	60%	70%	80%
Nex-CP WLS coverage	0.90	0.91	0.89	0.89
`EnbPI` coverage	0.88	0.88	0.87	0.86
`SPCI` coverage	0.89	0.90	0.89	0.90
Nex-CP WLS width	154.2	135.42	136.19	125.06
`EnbPI` width	110.01	103.32	106.56	100.31
`SPCI` width	70.95	77.33	67.19	70.66

(a) Original $X_t, \alpha = 0.1$.

Training fraction	50%	60%	70%	80%
Nex-CP WLS coverage	0.91	0.90	0.89	0.90
`EnbPI` coverage	0.89	0.88	0.88	0.90
`SPCI` coverage	0.89	0.90	0.89	0.90
Nex-CP WLS width	99.74	99.71	101.00	103.96
`EnbPI` width	73.20	54.18	52.90	50.37
`SPCI` width	58.83	48.55	42.24	44.61

(b) Considering non-stationarity, $\alpha = 0.1$.

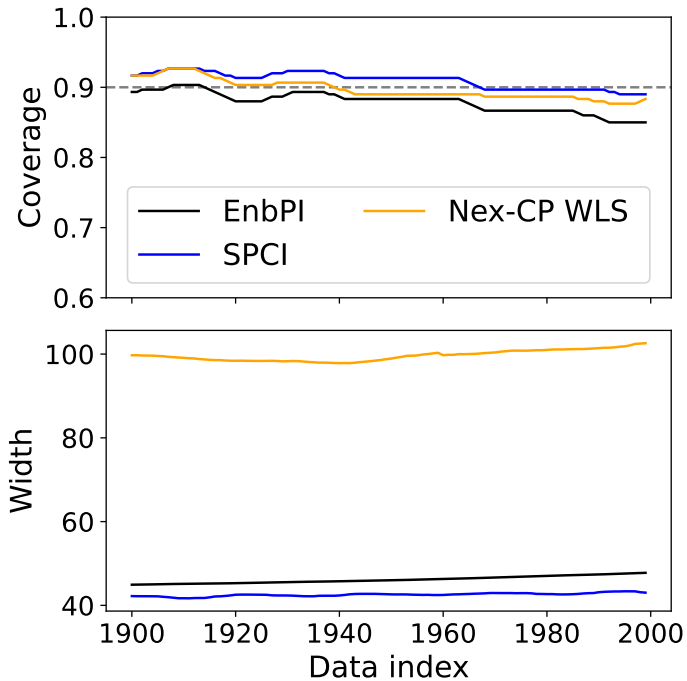


Figure: Rolling coverage considering non-stationarity; Training Fraction = 80% or 1600 obs.

Data in (Barber et al. 2022)

- Nex-CP WLS (Barber et al. 2022) vs. `EnbPI` and `SPCI`
- Target coverage is 0.9.
- Both `EnbPI` and `SPCI` yield narrower intervals.

Training fraction	50%	60%	70%	80%
Nex-CP WLS coverage	0.89	0.92	0.90	0.90
`EnbPI` coverage	0.88	0.89	0.88	0.91
`SPCI` coverage	0.92	0.92	0.91	0.93
Nex-CP WLS width	0.49	0.46	0.44	0.45
`EnbPI` width	0.42	0.39	0.36	0.32
`SPCI` width	0.23	0.22	0.22	0.22

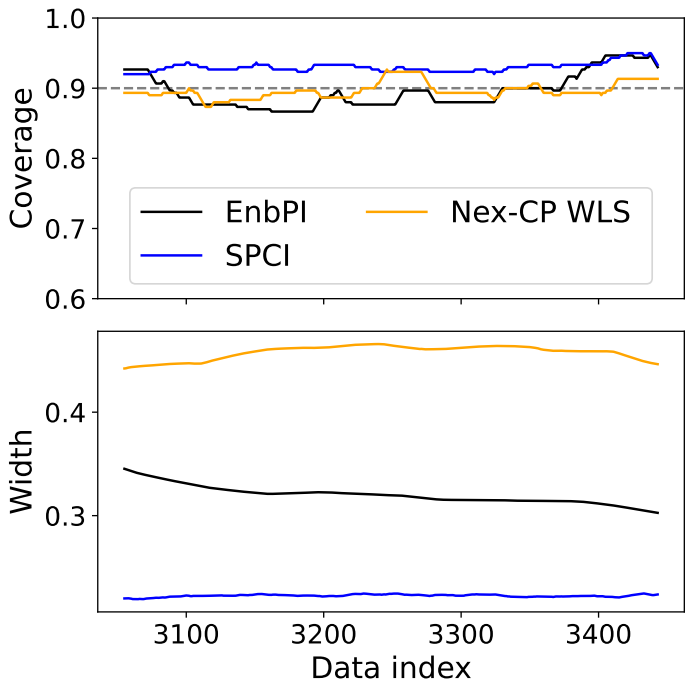
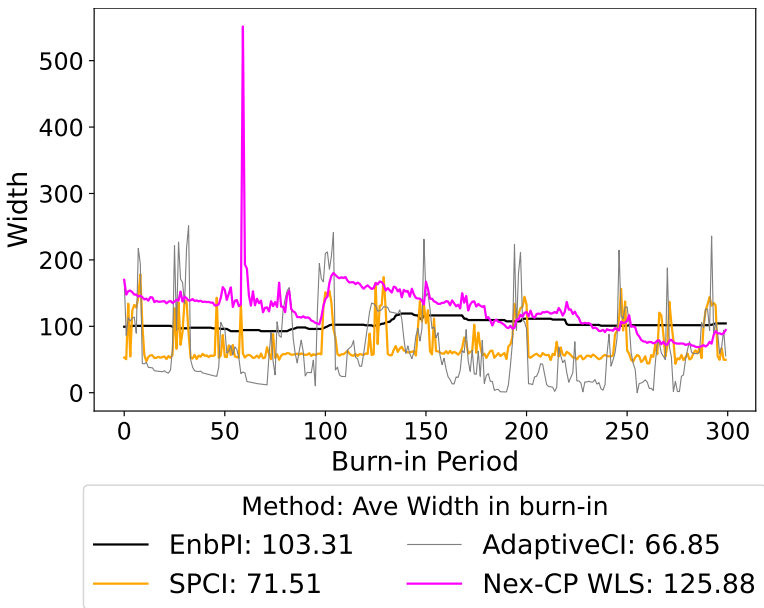


Figure: Rolling coverage; Training Fraction = 80%.

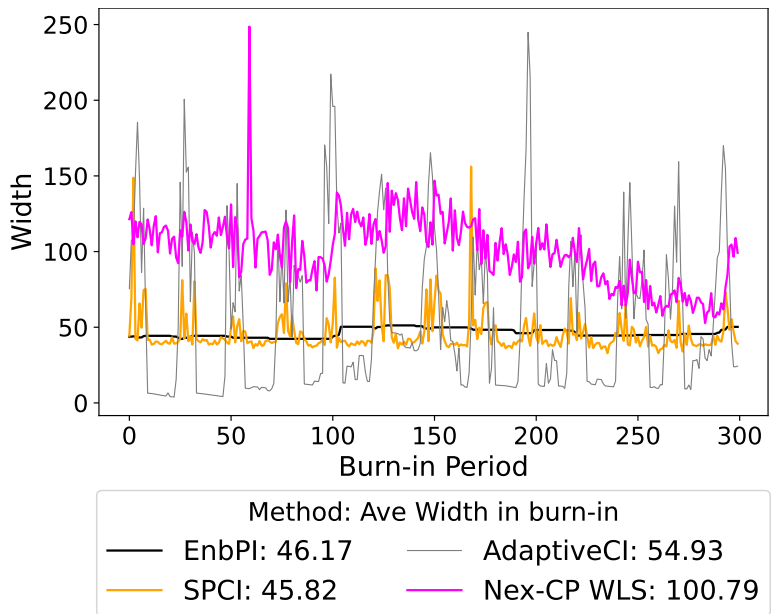
Back to solar data

Analyze burn-in period

- Figures above visualize average width over rolling windows.
- Examine intervals at "burn-in" period for $t = T + 1, T + 2, \ldots$
- (1) EnbPI: stable, not adaptive. (2) AdaptiveCI: high variance. (3) SPCI: adaptive, smaller variance. (4) Nex-CP: larger width.



(a) Original X_t.



(a) Considering non-stationarity.

Sequential conformal prediction set

- Aforementioned procedures apply to classification algorithms: response Y_t is categorical
- Coupling with the non-conformity score defined in (Angelopoulos et al., 2021), we call our method *Ensemble Regularized Adaptive Prediction Set* (ERAPS).
- SRAPS in (Angelopoulos et al., 2021), SAPS in (Romano et al., 2020), and Naive returns top-k labels whose cumulative probabilities exceed $1 - \alpha$.
- (Details can be found in the poster session)

α	0.05		0.075		0.1		0.15		0.2	
Pedestrain	coverage	set size	coverage	set size	coverage	set size	coverage	set size	coverage	set size
ERAPS	0.94	1.69	0.92	1.18	0.90	1.04	0.85	0.96	0.81	0.91
SRAPS	0.95	4.09	0.94	3.25	0.92	3.00	0.89	2.02	0.82	1.17
SAPS	0.95	4.29	0.93	3.77	0.91	3.00	0.86	2.16	0.81	1.86
Naive	0.87	1.60	0.84	1.47	0.81	1.37	0.75	1.22	0.71	1.10

Conformal prediction set for time-series, Xu, **X**. June 2022.
https://arxiv.org/abs/2206.07851

Summary

- `SPCI` and `EnbPI`: general conformal prediction for time series (non-exchangeable, temporally dependent, and non-stationary)
- `SPCI`
  - **narrower width** with a target coverage rate
  - exploiting **temporal dependence** and **feedback** of residuals
  - based on sequential quantile prediction for residuals
- Coverage guarantees
  - `EnbPI`: based on *data regularity* and *estimator quality*
  - `SPCI` (Ongoing): developing based on guarantees for time-series quantile estimation (Cai, 2002; Zhou and Wu, 2009; Biau and Patra, 2011; Xiao, 2012)
- `EnbPI` incorporated in `Scikit-learn`, `MAPIE`; Working with Meta to incorporate into Kats
- Ongoing: Extend to spatio-temporal (joint prediction of multi-dimensional time series)

Conformal prediction for time-series. Xu and **X**.
`https://arxiv.org/abs/2010.09107`

ICML 2021 (Long presentation).

Thank you!