Sequential Conformal Prediction for Time Series

Chen Xu, Yao Xie

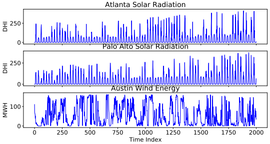
 

H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology

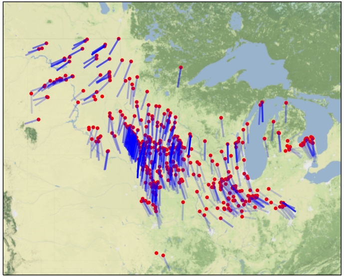
HKUST, Seminar on statistics and data science

Time series data

Solar energy prediction



Wind power prediction



Supply chain demand forecasting

ICU sequential data prediction

Time series prediction with uncertainty quantification

$$x_1, \ldots, x_t \longrightarrow \boxed{f_t(\cdot)} \longrightarrow y_t$$

- Quantify uncertainty of a chosen prediction algorithm f_t for any data?
- For applications (wind, solar, supply chain, medical) crucial to not only point predictor, but also given "confidence interval" as input to subsequent decision

nature > npj digital medicine > articles > article

Article | Open Access | Published: 09 September 2021

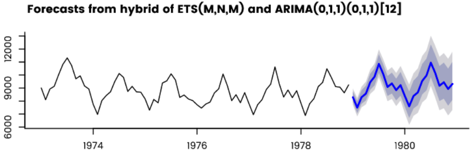
Artificial intelligence sepsis prediction algorithm learns to say "I don't know"

Supreeth P. Shashikumar, Gabriel Wardi, Atul Malhotra & Shamim Nemati

npj Digital Medicine **4**, Article number: 134 (2021) | Cite this article

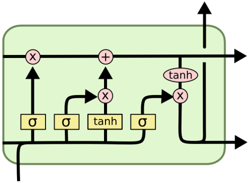
Prediction interval time series?

- Traditional time-series models (e.g. ARMA) has analytical prediction interval



Peter's stats stuff - R

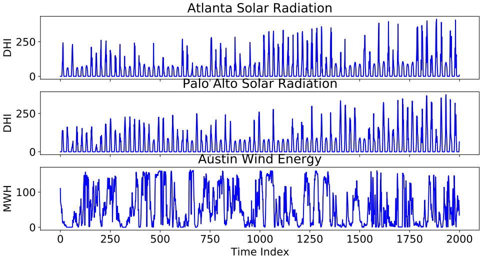
- Black-box machine learning models (e.g., RNN, LSTM), better performance for complex real data, but harder to come up with prediction interval with guarantees

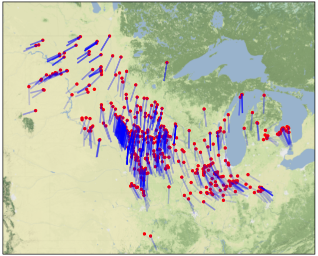


Peter's stats stuff - R

Conformal prediction for time series?

- Challenges for developing conformal prediction for time-series data
  - Consider non-stationary time series
  - Data are not exchangeable
  - Complex temporal correlation in data





Problem setup

- Constructing prediction intervals that attain valid coverage in finite samples, without making parametric distributional assumptions.

- Time series conformal prediction

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

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

- Also known as non-linear time series model in statistics (Fan and Yao 2003)

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

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

Goal

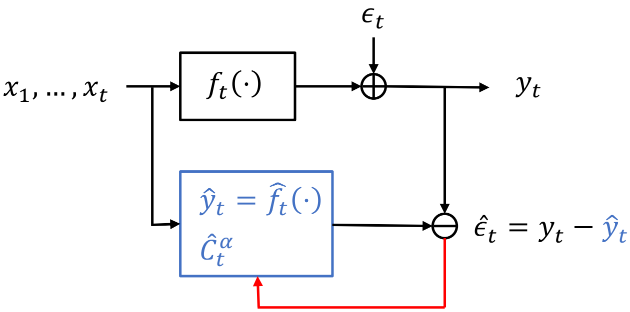
- Given a prediction algorithm $\hat{f}_t$ trained using data $\{x_t, y_t\}, t = 1, \ldots, T$, that generates predictions for $t = T+1, T+2, \ldots$
- Goal: Quantify the uncertainty of time series prediction algorithm $\hat{f}_t(X_t)$, $t > T$
- Construct prediction intervals $\widehat{C}_t^{\alpha}, t > T$, with pre-specified significance level $\alpha > 0$
  - *marginal* coverage guarantee:

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

  - *conditional* coverage guarantee

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

Sequential conformal inference for time-series



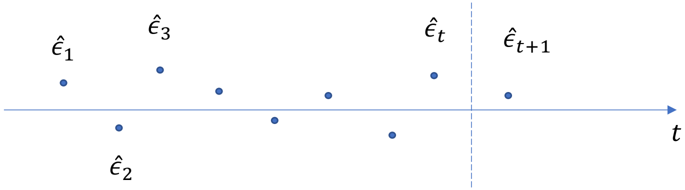
- Data not exchangeable
- Feedback available:
  Algorithm predicts $\hat{Y}_t \rightarrow$ True Y_t reveals $\rightarrow$ Feedback $\hat{\epsilon}_t$
- Nature can generate temporally correlated ϵ_t with unknown pdf

Sequential conformal inference

- Prediction algorithm $\hat{f}_t$ trained using past data
- Prediction residual

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

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

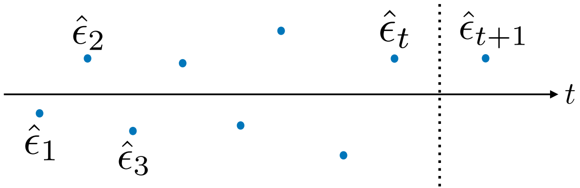


Conformal prediction for time-series. Xu, **X**. ICML 2021. (Long Talk)

Traditional vs. sequential conformal inference

$$\left(\hat{\epsilon}_1, \hat{\epsilon}_2, \ldots, \hat{\epsilon}_t\right) \overset{d}{=} \left(\hat{\epsilon}_{\sigma(1)}, \hat{\epsilon}_{\sigma(2)}, \ldots, \hat{\epsilon}_{\sigma(t)}\right)$$

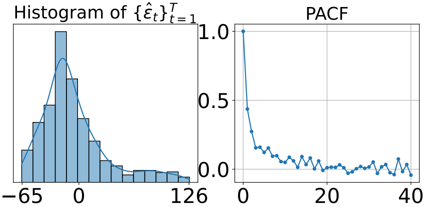
Traditional CP



Sequential CP

What do residuals $\hat{\epsilon}_t$ look like?

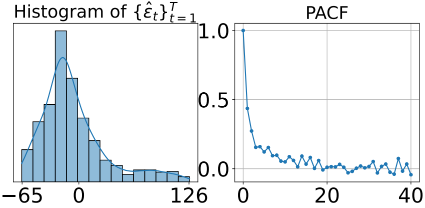
- Solar power radiation prediction for downtown Atlanta, Georgia
- Random forest for one-step-ahead prediction



- Asymmetric residual distribution
- Residuals have temporal correlation

What's in prediction residuals $\hat{\epsilon}_t$?

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



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

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

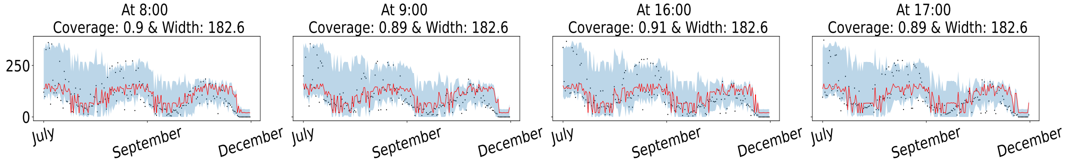
Sequential conformal inference

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

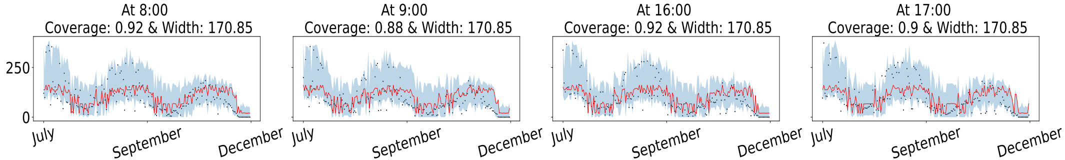
- Sequential Predictive Conformal Inference (SPCI):
  Exploiting temporal dependence of residuals
  - **Quantile regression** to get $\mathbb{P}\{\hat{\epsilon}_t > x | \hat{\epsilon}_{t-1}, \ldots, \hat{\epsilon}_{t-w}\}$
  - Guarantee for stationary residuals – allowing strong dependence

Solar power prediction

- Coverage: $\texttt{SPCI} \approx \texttt{EnbPI}$
- Interval width: $\texttt{SPCI} < \texttt{EnbPI}$



(a) EnbPI conditional coverage and width at each hour



(b) SPCI conditional coverage and width at each hour

Non-sequential conformal inference

Requires data exchangeability

- Split conformal (Vovk et al. 2005)

Training	holdout	X_{n+1}

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

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

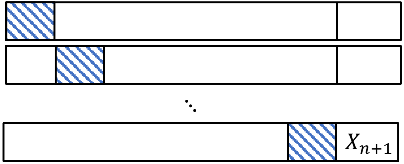
(Vovk et al. 2005), Lasso (Lei 2019)

Training	(X_{n+1}, y)

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

- 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 and time-series data

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

EnbPI

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

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

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

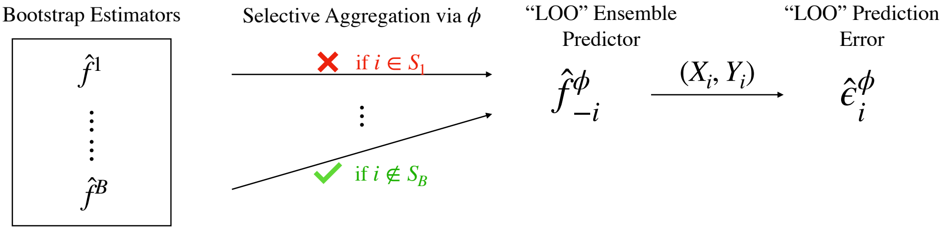
Q_α computes empirical α quantile of $\mathcal{E}_{t-1} := \{\hat{\epsilon}_i\}_{i=t-1,\ldots,t-w}$

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

Practical implementation

Ensemble Batch Prediction Interval (EnbPI) Algorithm

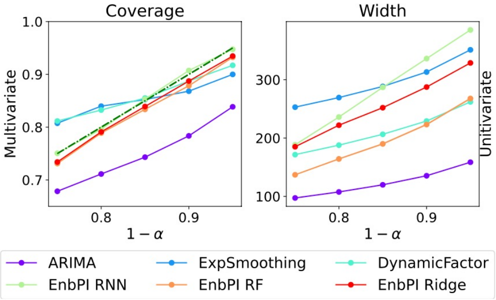
- Inspired by `J+aB`: Jackknife+-after-bootstrap (Kim, Xu, Barber 2020)
- In ensemble learning (e.g., bootstrap aggregation), multiple bootstrap models $\hat{f}^b$ are aggregated via ϕ (e.g., mean, median, weighted average) to improve prediction accuracy.
- Efficiently compute each $\hat{f}_{-t}$ using ensemble predictor



Example: Solar power prediction

Table 3: Solar power prediction in Atlanta, comparison of **EnbPI** with AdaptCI, ARIMA, Exponential Smoothing, and Dynamic Factor Models. We vary $\alpha \in [0.05, 0.10, 0.15, 0.20]$ and use the first 20% data as training data.

α	0.05					0.10					0.15					0.20				
Method	EnbPI	AdaptCI	ARIMA	Exp Smoothing	Dynamic Factor	EnbPI	AdaptCI	ARIMA	Exp Smoothing	Dynamic Factor	EnbPI	AdaptCI	ARIMA	Exp Smoothing	Dynamic Factor	EnbPI	AdaptCI	ARIMA	Exp Smoothing	Dynamic Factor
Coverage	0.950	0.863	0.839	0.900	0.917	0.896	0.831	0.784	0.868	0.887	0.846	0.806	0.743	0.852	0.855	0.798	0.776	0.711	0.840	0.832
Width	288.581	215.258	158.581	351.181	262.006	216.989	187.504	135.404	313.185	229.151	178.140	173.079	119.870	288.428	206.448	147.297	154.322	107.652	269.379	187.840



Further improving EnbPI?

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

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

- EnbPI is based on empirical distribution of $\{\hat{\epsilon}_t\}$
- What's typical characteristic to time-series (*of residuals*)?

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

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

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

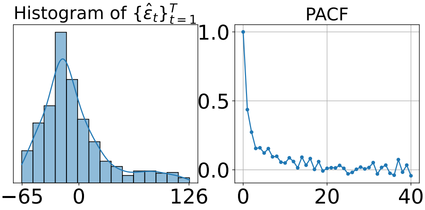
Further exploiting temporal dependence: `SPCI`

`Sequential Predictive Conformal Inference (SPCI)`

- Idea: Estimate $\widehat{C}_t^\alpha$ by predicting residual quantile from past observed residuals:

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

- Use *quantile regression* (e.g., random forest (Meinshausen 2006), nearest-neighbor based (Biau & Patra 2011)) on residuals (conformity scores)

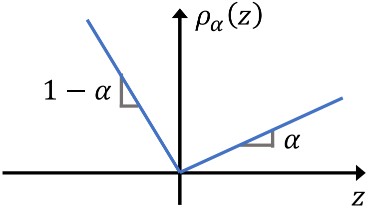


Quantile regression

- 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



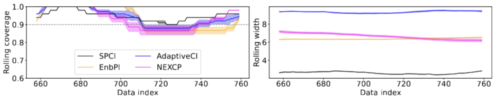
Comparison with other time-series methods

Table 3: Marginal coverage and width by all methods on three real time series. The target coverage is 0.9, and entries in the bracket indicate standard deviation over three independent trials. `SPCI` outperforms competitors with a much narrower interval width and does not lose coverage.

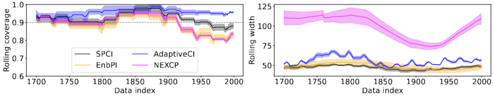
	Wind coverage	Wind width	Electric coverage	Electric width	Solar coverage	Solar width
SPCI	0.95 (1.50e-02)	2.65 (1.60e-02)	0.93 (4.79e-03)	0.22 (1.68e-03)	0.91 (1.12e-02)	47.61 (1.33e+00)
EnbPI	0.93 (6.20e-03)	6.38 (3.01e-02)	0.91 (6.84e-04)	0.32 (9.11e-04)	0.88 (4.25e-03)	48.95 (3.38e+00)
AdaptiveCI	0.95 (5.37e-03)	9.34 (3.56e-02)	0.95 (1.81e-03)	0.51 (7.25e-03)	0.96 (1.39e-02)	56.34 (1.15e+00)
NEX-CP	0.96 (8.21e-03)	6.68 (7.73e-02)	0.90 (2.05e-03)	0.45 (2.16e-03)	0.90 (7.73e-03)	102.80 (5.25e+00)

The ELEC2 data set4 [Harries, 1999] tracks electricity usage and pricing in the states of New South Wales and Victoria in Australia, every 30 minutes over a 2.5 year period in 1996–1999.

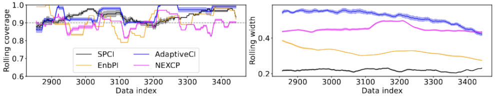
Comparison with other time-series conformal prediction



(a) Wind



(b) Solar



(c) Electric

Figure 4: Rolling coverage and interval width over three real time series by different methods. `SPCI` in black not only yields valid rolling coverage but also consistently yields the narrowest prediction intervals. Furthermore, the variance of `SPCI` results over trials is also small, as shown by the shaded regions over coverage and width results.

Theoretical guarantee: `EnbPI`

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

Consider $f_t(X_t) = f(X_t)$:

- 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)

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

Theoretical guarantee (cont.)

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

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

<u>Implications</u>

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

Assumption 1 can be extended

- Independent $\{\epsilon_t\}_{t\geq 1}$

$$\text{Rate} = (\log(16T)/T)^{1/2}.$$

- Stationary linear processes $\epsilon_t = \sum_{j=1}^{\infty} \delta_j z_{t-j}$.

$$\text{Rate} = \log T/\sqrt{T}$$

  Faster than strongly mixing errors, slower than independent errors.

- Joint density of $\{\epsilon_t\}_{t=1}^{T+1}$ satisfies a logarithmic Sobolev inequality

$$\text{Rate} = (\log(cT)/T)^{1/3}$$

Assumption 2: "Good" predictive algorithm

- Assumption 2 holds true for many classes of algorithms
- No-free-lunch theorem:
  assumption on f is necessary in order for us to approximate it well.
- Examples
  - if f is sufficiently smooth,
    $$\delta_T = o(T^{-1/4})$$
    for neural networks sieve estimators (Chen and White, 1999).
  - If f is a sparse high-dimensional linear model,
    $$\delta_T = o(T^{-1/2})$$
    for Lasso and Dantzig selector (Bickel et al. 2009).

Summary

- Sequential conformal prediction for time series (non-exchangeable, temporally dependent, and non-stationary)
- Two algorithms `EnbPI` and `SPCI`
  - `EnbPI` based on empirical residuals
  - `SPCI` exploiting **temporal dependence** of residuals
- Handling non-stationarity and heteroskedasticity
- Our algorithms are incorporated in `Scikit-learn/MAPIE`; AWS Fortuna Time Series Package; Meta to incorporate into Kats.
- Can be generalized to sequential conformal prediction set, anomaly detection

Conformal prediction for time-series. Xu and **X**. ICML 2021 (Long Talk). IEEE TPAMI 2023.
Sequential Predictive Conformal Inference for Time Series. Xu and **X**. ICML 2023.
Conformal prediction set for time-series, Xu, **X**. June 2022. https://arxiv.org/abs/2206.07851

Thank NSF CAREER CCF-1650913, and NSF DMS-2134037, CMMI-2015787, CMMI-2112533, DMS-1938106, and DMS-1830210.