Conformal Prediction for Time Series

An application to forecasting French electricity Spot prices

Margaux Zaffran[1,2,3] Aymeric Dieuleveut[3] Olivier Féron[1,4] Yannig Goude[1] Julie Josse[2]

15/12/2021

[1]EDF R&D [2]INRIA [3]CMAP, Ecole Polytechnique [4]FiME

Forecasting French electricity Spot prices

Electricity Spot prices

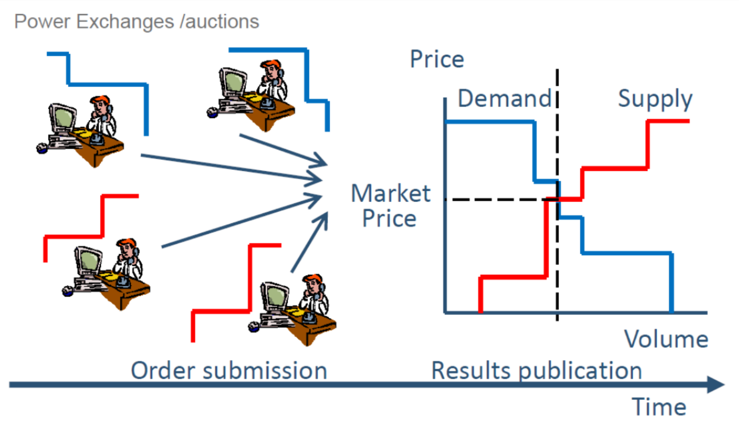


Figure 1: Drawing of spot auctions mechanism

French Electricity Spot prices data set: visualisation

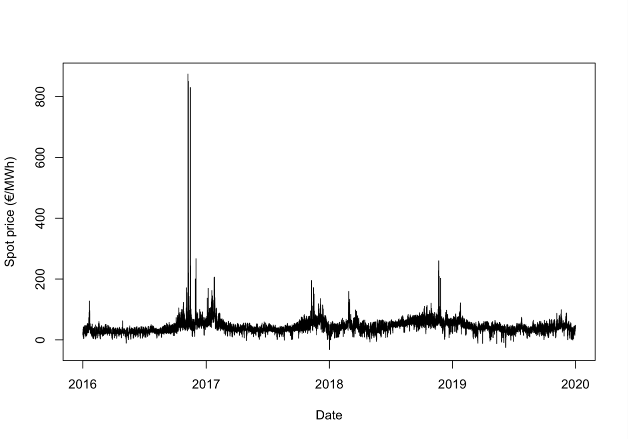


Figure 2: Representation of the French electricity spot price, from 2016 to 2019.

French Electricity Spot prices data set: extract

Date and time	Price	Price D-1	Price D-7	For. cons.	DOW
11/01/16 0PM	21.95	15.58	13.78	58800	Monday
11/01/16 1PM	20.04	19.05	13.44	57600	Monday
⋮	⋮	⋮	⋮	⋮	⋮
12/01/16 0PM	21.51	21.95	25.03	61600	Tuesday
12/01/16 1PM	19.81	20.04	24.42	59800	Tuesday
⋮	⋮	⋮	⋮	⋮	⋮
18/01/16 0PM	38.14	37.86	21.95	70400	Monday
18/01/16 1PM	35.66	34.60	20.04	69500	Monday
⋮	⋮	⋮	⋮	⋮	⋮

Table 1: Extract of the built data set, for French electricity spot price forecasting.

Forecasting French electricity Spot prices

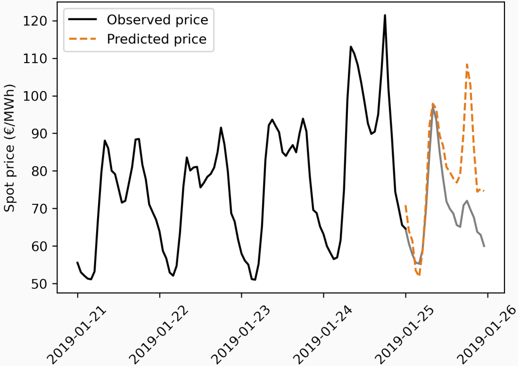


Figure 3: French electricity spot price and its prediction with random forest.

$\hookrightarrow$ $(x_t, y_t) \in \mathbb{R}^d \times \mathbb{R}$ ($d = 56$, details later)

$\hookrightarrow$ 3 years for training

$\hookrightarrow$ 1 year to forecast

Forecasting French electricity Spot prices with confidence

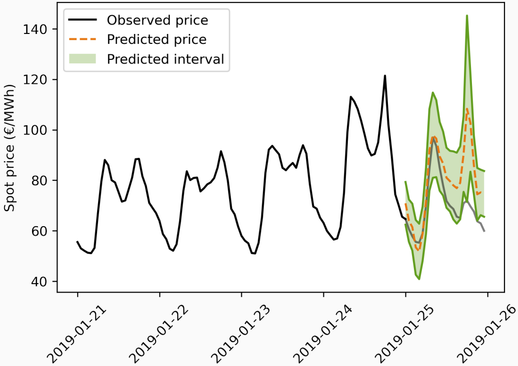


Figure 4: French electricity spot price, its prediction and its uncertainty with Adaptive Conformal Inference (Gibbs and Candès, 2021).

Forecasting French electricity Spot prices with confidence: results

- Target coverage: 90%
- Empirical coverage: 90.46%[1]
- Average length: 22.91€/MWh

[1] But conditional coverage varies from 86.14% to 93% depending on the day of the week (from example).

Available methods for non-exchangeable data, in the context of time series

Framework and notations

- Data: T_0 observations $(x_1, y_1), \ldots, (x_{T_0}, y_{T_0})$ in $\mathbb{R}^d \times \mathbb{R}$
- Aim: predict the response values as well as predictive intervals for T_1 subsequent observations $x_{T_0+1}, \ldots, x_{T_0+T_1}$

$\hookrightarrow$ Build the smallest interval $\mathcal{C}_\alpha^t$ such that:

$$\mathbb{P}\left\{ Y_t \in \mathcal{C}_\alpha^t\left(X_t\right)\right\} \geq 1 - \alpha, \text{ for } t \in [\![T_0 + 1; T_0 + T_1]\!].$$

How to adapt to time series?

Usual ideas from the time series literature:

- Consider an online procedure (for each new data, re-train and re-calibrate)
  - $\hookrightarrow$ update to recent observations (trend impact, period of the seasonality, dependence...)

How to adapt to time series?

Usual ideas from the time series literature:

- Consider an online procedure (for each new data, re-train and re-calibrate)
  - $\hookrightarrow$ update to recent observations (trend impact, period of the seasonality, dependence...)
- Use a sequential split
  - $\hookrightarrow$ use only the past so as to correctly estimate the variance of the residuals (using the future leads to optimistic residuals and underestimation of their variance)

Available methods

- Online (sequential) split conformal prediction (Wisniewski et al. (2020); Kath and Ziel (2021); and our study);
  - $\hookrightarrow$ tested on real time series

Available methods

- Online (sequential) split conformal prediction (Wisniewski et al. (2020); Kath and Ziel (2021); and our study);
  - $\hookrightarrow$ tested on real time series
- Ensemble Prediction Interval (Xu and Xie, 2021);
  - $\hookrightarrow$ tested on other real time series
  - $\hookrightarrow$ compared to offline methods (unfair)

Available methods

- Online (sequential) split conformal prediction (Wisniewski et al. (2020); Kath and Ziel (2021); and our study);
  - $\hookrightarrow$ tested on real time series
- Ensemble Prediction Interval (Xu and Xie, 2021);
  - $\hookrightarrow$ tested on other real time series
  - $\hookrightarrow$ compared to offline methods (unfair)
- Adaptive Conformal Inference (Gibbs and Candès, 2021).
  - $\hookrightarrow$ tested on one simulation and real time series with important breaks (distribution shift)

Available methods

- Online (sequential) split conformal prediction (Wisniewski et al. (2020); Kath and Ziel (2021); and our study);
  - $\hookrightarrow$ tested on real time series
- Ensemble Prediction Interval (Xu and Xie, 2021);
  - $\hookrightarrow$ tested on other real time series
  - $\hookrightarrow$ compared to offline methods (unfair)
- Adaptive Conformal Inference (Gibbs and Candès, 2021).
  - $\hookrightarrow$ tested on one simulation and real time series with important breaks (distribution shift)

$\Rightarrow$ No systematic simulations

$\Rightarrow$ No fair and common comparison

Online sequential conformal prediction (OSCP)

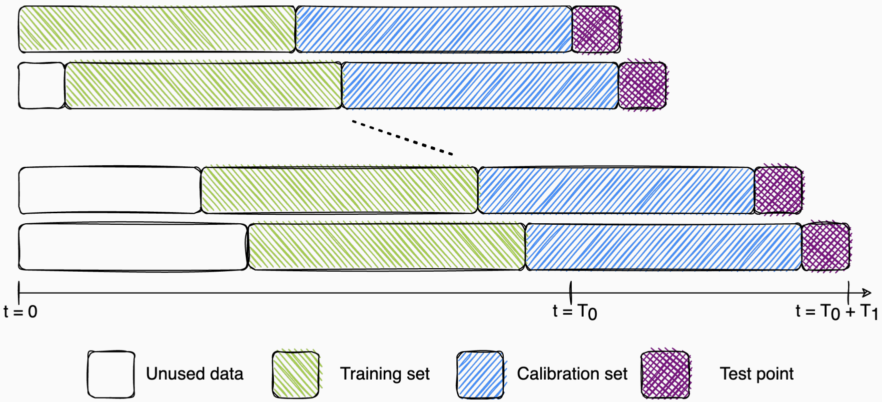


Figure 5: Diagram describing the online sequential split conformal prediction.

EnbPI, Xu and Xie (2021)

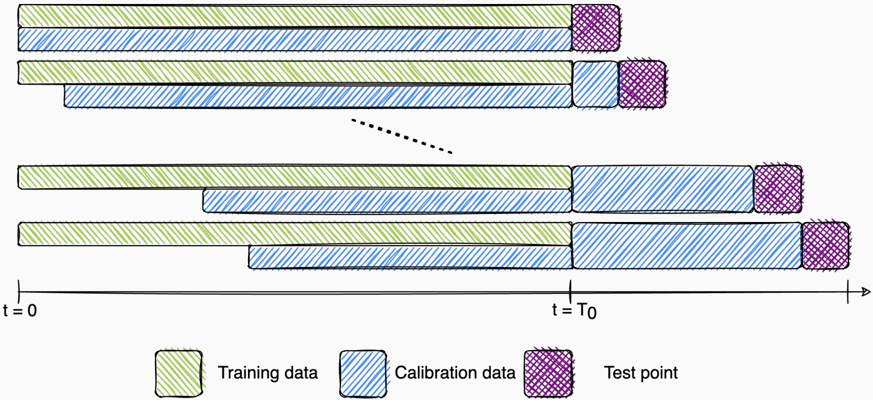
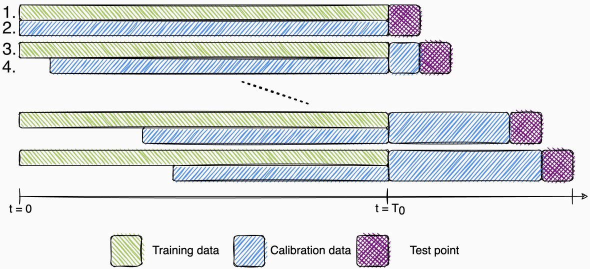


Figure 6: Diagram describing the EnbPI algorithm.

EnbPI, Xu and Xie (2021)



1. Train B bootstrap predictors;
2. Obtain out-of-bootstrap residuals by aggregating the corresponding predictors;
3. Do not re-train the B bootstrap predictors;
4. Obtain new residual by aggregating all the predictors. Forget the first residuals.

Adaptive Conformal Inference (ACI), Gibbs and Candès (2021)

Refitting the model may be insufficient $\Rightarrow$ adapt the quantile level used on the calibration's scores.

Adaptive Conformal Inference (ACI), Gibbs and Candès (2021)

Refitting the model may be insufficient $\Rightarrow$ adapt the quantile level used on the calibration's scores. **(Distribution shift)**

Adaptive Conformal Inference (ACI), Gibbs and Candès (2021)

Refitting the model may be insufficient $\Rightarrow$ adapt the quantile level used on the calibration's scores. **(Distribution shift)**

The proposed update scheme is the following:

$$\alpha_{t+1} := \alpha_t + \gamma\left(\alpha - \mathrm{err}_t\right) \tag{1}$$

with:

$$\mathrm{err}_t := \begin{cases} 1, & \text{if } y_t \notin \hat{\mathcal{C}}_{\alpha_t}\left(x_t\right), \\ 0, & \text{otherwise}, \end{cases}$$

and $\alpha_1 = \alpha$, $\gamma \geq 0$.

Adaptive Conformal Inference (ACI), Gibbs and Candès (2021)

Refitting the model may be insufficient $\Rightarrow$ adapt the quantile level used on the calibration's scores. **(Distribution shift)**

The proposed update scheme is the following:

$$\alpha_{t+1} := \alpha_t + \gamma\left(\alpha - \mathrm{err}_t\right) \tag{1}$$

with:

$$\mathrm{err}_t := \begin{cases} 1, & \text{if } y_t \notin \hat{\mathcal{C}}_{\alpha_t}\left(x_t\right), \\ 0, & \text{otherwise }, \end{cases}$$

and $\alpha_1 = \alpha$, $\gamma \geq 0$.

Intuition: if we did make an error, the interval was too small so we want to increase its length by taking a higher quantile (a smaller α_t). Reversely if we included the point.

Visualisation of the procedure

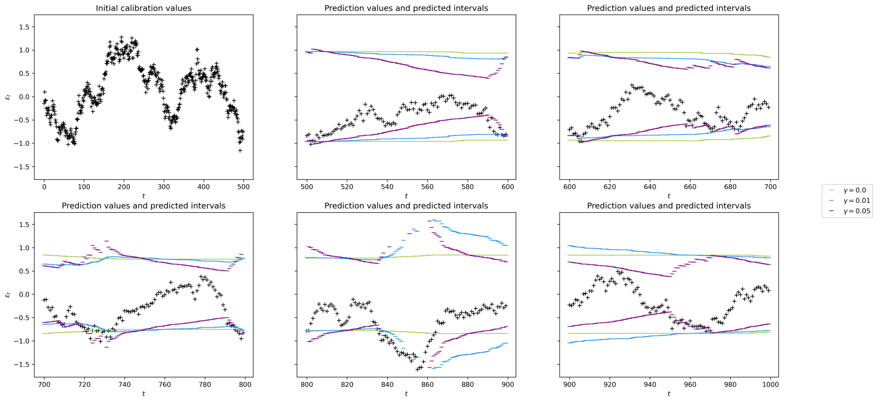


Figure 6: Visualisation of ACI with different values of γ

Visualisation of the procedure

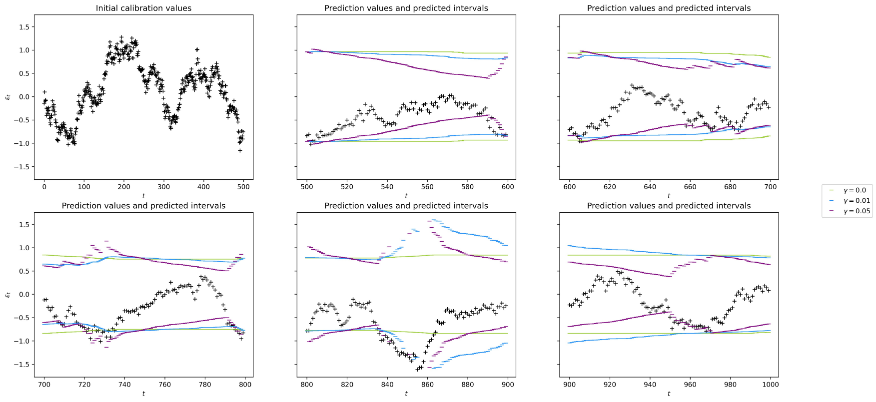


Figure 6: Visualisation of ACI with different values of γ

ACI originally splitted randomly. We use ACI with a **sequential split**.

Summary of the methods

Methods	Pros	Cons
OSCP	• Easy to implement	• No general theoretical validity (results hold until strongly mixing[2])
EnbPI	• Adapted to small data sets • Quicker on new forecasts	• Bootstrap not adapted to time series • Mixes two different aggregation functions[3] • No general theoretical validity
ACI	• Easy to implement • Theoretical validity without assumptions (long-term)	• γ tuning

[2] Chernozhukov et al. (2018)

[3] New paper changing this, after discussion with Chen Xu at ICML workshop.

Computational aspect

	Currently available		Contribution	
Methods	Language	Details	Language	Options
CP	R		Python	
OSCP	not available		Python	randomized split
EnbPI	Python		Python	same aggregation function
ACI	R script	no general function	Python	randomized split

$\Rightarrow$ We propose a unified repository containing all the conformal prediction methods for time series, with their variants as options.

Comparison on simulated data

Data generation

$$Y_t = 10\sin\left(\pi X_{t,1}X_{t,2}\right) + 20\left(X_{t,3} - 0.5\right)^2 + 10X_{t,4} + 5X_{t,5} + \varepsilon_t$$

where the X_t are multivariate uniformly distributed on $[0,1]$ and ε_t are generated from an ARMA(1,1) process.

Data generation

$$Y_t = 10\sin(\pi X_{t,1}X_{t,2}) + 20(X_{t,3}-0.5)^2 + 10X_{t,4} + 5X_{t,5} + \varepsilon_t$$

where the X_t are multivariate uniformly distributed on $[0,1]$ and ε_t are generated from an ARMA(1,1) process.

Definition (ARMA(1,1) process)

We say that ε_t is an ARMA(1,1) process if for any t:

$$\varepsilon_{t+1} = \varphi\varepsilon_t + \xi_{t+1} + \theta\xi_t,$$

with ξ_t is a white noise of variance σ^2, called the **innovation**.

Data generation

$$Y_t = 10\sin\left(\pi X_{t,1}X_{t,2}\right) + 20\left(X_{t,3} - 0.5\right)^2 + 10X_{t,4} + 5X_{t,5} + \varepsilon_t$$

where the X_t are multivariate uniformly distributed on $[0,1]$ and ε_t are generated from an ARMA(1,1) process.

Definition (ARMA(1,1) process)

We say that ε_t is an ARMA(1,1) process if for any t:

$$\varepsilon_{t+1} = \varphi\varepsilon_t + \xi_{t+1} + \theta\xi_t,$$

with ξ_t is a white noise of variance σ^2, called the **innovation**.

- $\varphi = \theta$ range in $[0.1, 0.8, 0.9, 0.95, 0.99]$.
- We fix σ so as to keep the variance $\mathrm{Var}(\varepsilon_t)$ constant to 1 or 10.

Simulation settings

We use random forest as regressor.

Simulation settings

We use random forest as regressor.

For each setting (pair variance and φ,θ):

- 300 points, the last 100 kept for prediction and evaluation,
- 500 repetitions,

$\Rightarrow$ in total, $100 \times 500 = 50000$ predictions are evaluated.

We use random forest as regressor.

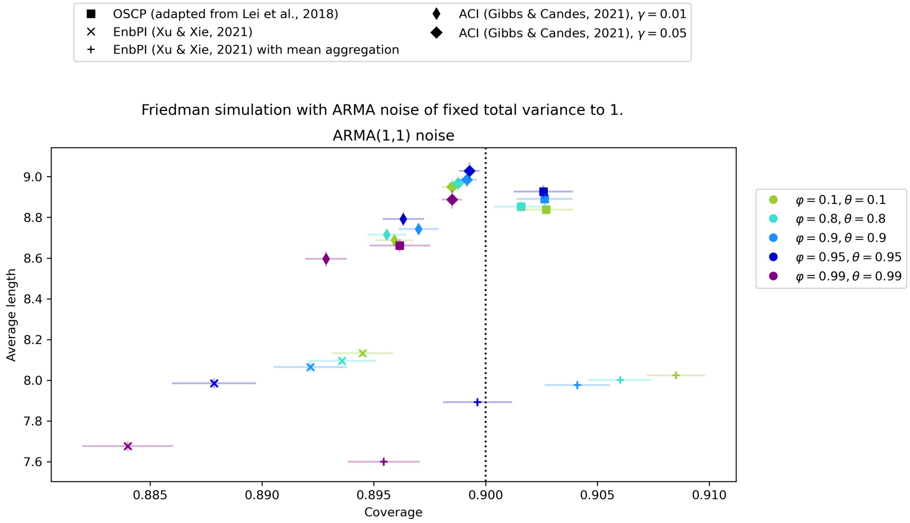
For each setting (pair variance and φ,θ):

- 300 points, the last 100 kept for prediction and evaluation,
- 500 repetitions,

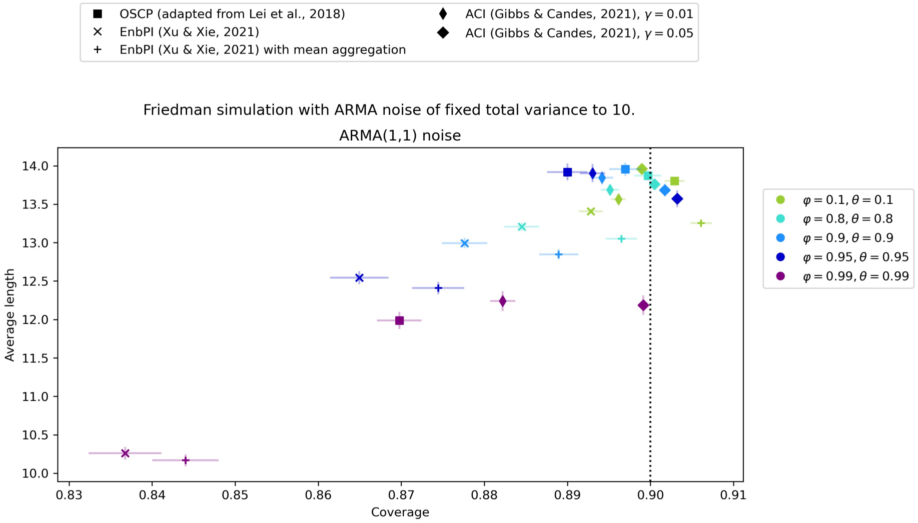
$\Rightarrow$ in total, $100 \times 500 = 50000$ predictions are evaluated.

We present the results in the ARMA(1,1) case, but we also have them for AR(1) and MA(1) processes.

Results: impact of the temporal dependence, variance 1



Results: impact of the temporal dependence, variance 10



Summary

- Online CP: achieves valid coverage for values of φ and θ smaller than 0.99.

Summary

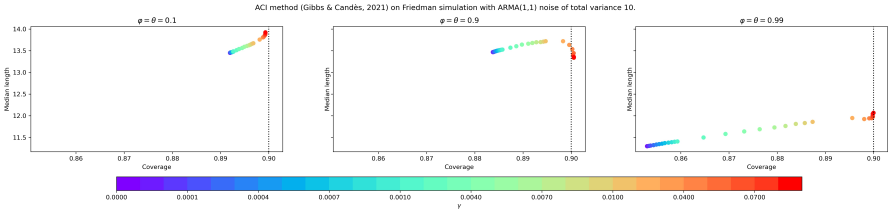
- Online CP: achieves valid coverage for values of φ and θ smaller than 0.99.
- ACI: achieves valid coverage with $\gamma = 0.05$. Nevertheless, the choice of γ is important.

Summary

- Online CP: achieves valid coverage for values of φ and θ smaller than 0.99.
- ACI: achieves valid coverage with $\gamma = 0.05$. Nevertheless, the choice of γ is important.
- EnbPI: for small variance, really competitive (small lengths). But for strong dependence and/or high variance, fails to attain coverage.

A closer look at ACI: choosing γ?

Empirical evaluation of ACI sensitivity to γ



$\Rightarrow$ The more the dependence, the more sensitive to γ is ACI.

Adaptive choice of γ

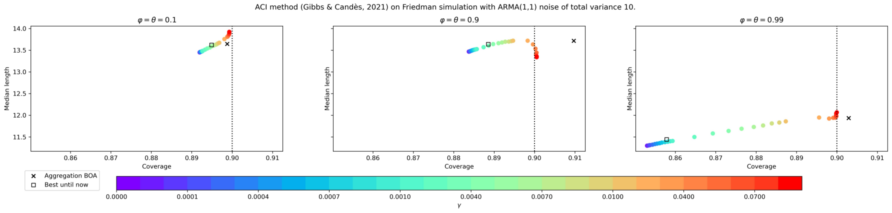
- Naive method: best until now (smallest among valid ones)

Adaptive choice of γ

- Naive method: best until now (smallest among valid ones)
- Improved method: online aggregation for each bound separately, using the pinball loss

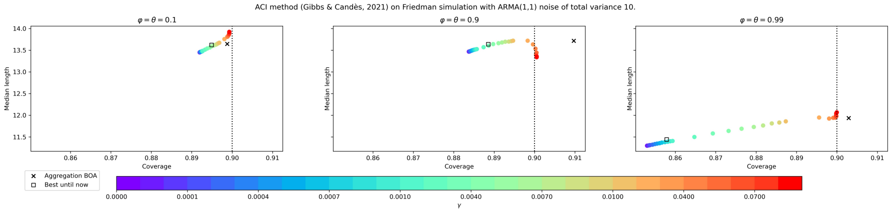
Adaptive choice of γ

- Naive method: best until now (smallest among valid ones)
- Improved method: online aggregation for each bound separately, using the pinball loss



Adaptive choice of γ

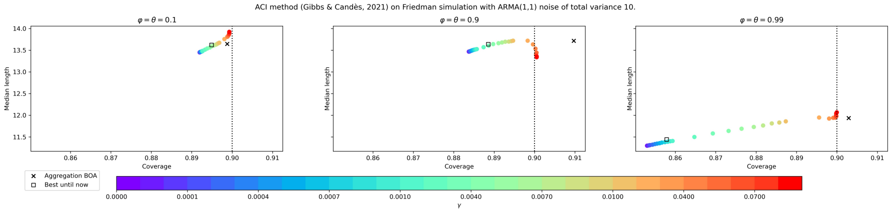
- Naive method: best until now (smallest among valid ones)
- Improved method: online aggregation for each bound separately, using the pinball loss



- Naive method: accumulates error of the different ACI's versions.

Adaptive choice of γ

- Naive method: best until now (smallest among valid ones)
- Improved method: online aggregation for each bound separately, using the pinball loss



- Naive method: accumulates error of the different ACI's versions.
- Expert aggregation: encouraging preliminary results.

Theoretical analysis of ACI's length

Aim: derive theoretical results on the **average length** of ACI depending on γ

$\hookrightarrow$ Guideline for choosing γ

Theoretical analysis of ACI's length

Aim: derive theoretical results on the **average length** of ACI depending on γ

$\hookrightarrow$ Guideline for choosing γ

Approach: consider extreme cases (useful in an adversarial context) even if strong assumptions are needed

1. i.i.d.
2. AR(1)
3. distribution shift
4. Hidden Markov Model

Theoretical analysis of ACI's length: i.i.d. case

Lemma

Assume that:

- $\alpha \in \mathbb{Q}$;
- *the scores are i.i.d. of quantile function* Q;
- *the quantile function is permanently perfectly estimated (i.e.* $\hat{Q}_t = Q$ *for all* $t > 0$*).*

Then $(\alpha_t)_t$ *forms an irreducible Markov Chain on a finite state space. Thus, it is a positive recurrent Markov Chain.*

Theoretical analysis of ACI's length: i.i.d. case

Theorem

Under the assumptions of previous lemma and that the quantile function Q is bounded.

Then we have:

$$\frac{1}{T}\sum_{t=1}^{T} L(\alpha_t) \underset{T\to+\infty}{\longrightarrow} \mathbb{E}_{\pi_\gamma}[L(\alpha_t)]$$

with π_γ the stationary distribution of the Markov Chain and:

$$\mathbb{E}_{\pi_\gamma}[L(\alpha_t)] = L_0 + \frac{Q''(1-\alpha)}{2}\gamma\alpha(1-\alpha) + O(\gamma^{3/2})$$

where:

- $L(\alpha_t) = 2Q(1-\alpha_t)$ *is the length of the adaptive algorithm (the dependence in γ is hidden in α_t, and $\gamma > 0$);*
- $L_0 = 2Q(1-\alpha)$ *is the length of the non-adaptive algorithm ($\gamma = 0$).*

Theoretical analysis of ACI's length: perspectives

- Similar results in the case where the scores are an AR(1) process
  - $\hookrightarrow$ exhibit an optimal γ depending on φ?

Theoretical analysis of ACI's length: perspectives

- Similar results in the case where the scores are an AR(1) process
  - $\hookrightarrow$ exhibit an optimal γ depending on φ?
- Similar results in the case where there is a distribution shift in the scores
  - $\hookrightarrow$ highlights the positive gain made by ACI

Theoretical analysis of ACI's length: perspectives

- Similar results in the case where the scores are an AR(1) process
  - $\hookrightarrow$ exhibit an optimal γ depending on φ?
- Similar results in the case where there is a distribution shift in the scores
  - $\hookrightarrow$ highlights the positive gain made by ACI
- Similar results in the case where there is a Hidden Markov Model

Price prediction with confidence in 2019

Settings

- Forecast for the year 2019.
- Random forest regressor.
- One model per hour, we concatenate the predictions afterwards.

Settings

- Forecast for the year 2019.
- Random forest regressor.
- One model per hour, we concatenate the predictions afterwards.

$\hookrightarrow$ 24 models

Settings

- Forecast for the year 2019.
- Random forest regressor.
- One model per hour, we concatenate the predictions afterwards.

$\hookrightarrow$ 24 models

  - $y_t \in \mathbb{R}$

Settings

- Forecast for the year 2019.
- Random forest regressor.
- One model per hour, we concatenate the predictions afterwards.

$\hookrightarrow$ 24 models

- $y_t \in \mathbb{R}$
- $x_t \in \mathbb{R}^d$, with $d = 24 + 24 + 1 + 7 = 56$

24 prices of the day before

Settings

- Forecast for the year 2019.
- Random forest regressor.
- One model per hour, we concatenate the predictions afterwards.

$\hookrightarrow$ 24 models

  - $y_t \in \mathbb{R}$
  - $x_t \in \mathbb{R}^d$, with $d = 24 + 24 + 1 + 7 = 56$

24 prices of the day before

24 prices of the 7 days before

Settings

- Forecast for the year 2019.
- Random forest regressor.
- One model per hour, we concatenate the predictions afterwards.

$\hookrightarrow$ 24 models

- $y_t \in \mathbb{R}$
- $x_t \in \mathbb{R}^d$, with $d = 24 + 24 + 1 + 7 = 56$

24 prices of the day before

24 prices of the 7 days before

Forecasted consumption

Settings

- Forecast for the year 2019.
- Random forest regressor.
- One model per hour, we concatenate the predictions afterwards.

$\hookrightarrow$ 24 models

  - $y_t \in \mathbb{R}$
  - $x_t \in \mathbb{R}^d$, with $d = 24 + 24 + 1 + 7 = 56$

24 prices of the day before

24 prices of the 7 days before

Forecasted consumption

Encoded day of the week

Settings

- Forecast for the year 2019.
- Random forest regressor.
- One model per hour, we concatenate the predictions afterwards.

$\hookrightarrow$ 24 models

  - $y_t \in \mathbb{R}$
  - $x_t \in \mathbb{R}^d$, with $d = \ 24 \ + \ 24 \ + \ 1 \ + \ 7 \ = 56$
  - 3 years for training/calibration, i.e. $T_0 = 1096$ observations

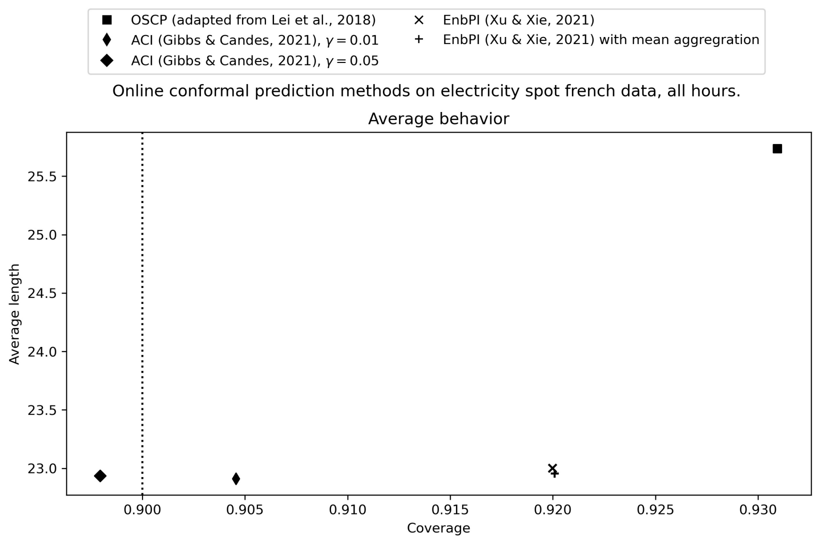
Settings

- Forecast for the year 2019.
- Random forest regressor.
- One model per hour, we concatenate the predictions afterwards.

$\hookrightarrow$ 24 models

  - $y_t \in \mathbb{R}$
  - $x_t \in \mathbb{R}^d$, with $d = \ 24 \ + \ 24 \ + \ 1 \ + \ 7 \ = 56$
  - 3 years for training/calibration, i.e. $T_0 = 1096$ observations
  - 1 year to forecast, i.e. $T_1 = 365$ observations

Performance on predicted French electricity Spot price for the year 2019



Performance on predicted French electricity Spot price: visualisation of a day

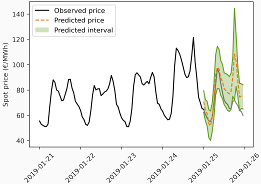


Figure 7: Online seq. split CP

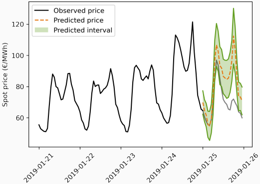


Figure 8: EnbPI

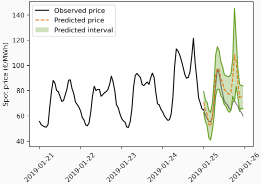


Figure 9: ACI with $\gamma = 0.01$

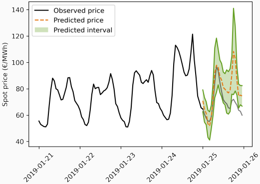


Figure 10: ACI with $\gamma = 0.05$

Perspective: towards conditional coverage?

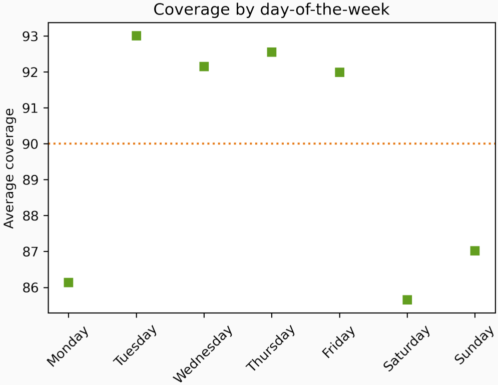


Figure 11: ACI with $\gamma = 0.05$

Concluding remarks

Conclusion

- Online sequential split conformal prediction achieves correct performances
- ACI obtains valid coverage in the time dependent settings, whilst designed initially for shifts
- ACI is sensitive to γ choice
- EnbPI is highly competitive in some regimes, but its performance depends a lot on the regime

Contributions

- Pipeline of analysis for simulation of increasing difficulty and real data analysis (code in python) for reproducible work and benchmarking conformal predictions in the framework of time series
- Demonstration of ACI's interest in the broader time series framework (simulation and real world)
- Theoretical results on ACI's length depending on γ (on-going)
- Empirical proposition of an adaptive choice of γ

Perspectives

- Refined analysis of expert aggregation for γ choice
  - Theoretical guarantees?
  - Other aggregation methods, other losses...

Perspectives

- Refined analysis of expert aggregation for γ choice
    - Theoretical guarantees?
    - Other aggregation methods, other losses...
- Development of a conformal prediction procedure for time series with approximate/asymptotic conditional coverage

Perspectives

- Refined analysis of expert aggregation for γ choice
    - Theoretical guarantees?
    - Other aggregation methods, other losses...
- **Development of a conformal prediction procedure for time series with approximate/asymptotic conditional coverage**

Perspectives

- Refined analysis of expert aggregation for γ choice
  - Theoretical guarantees?
  - Other aggregation methods, other losses...
- **Development of a conformal prediction procedure for time series with approximate/asymptotic conditional coverage**

  $\hookrightarrow$ ACI with $\alpha_t(x)$ and $\mathrm{err}_t(x)$?

Thank you!

References i

Chernozhukov, V., Wüthrich, K., and Yinchu, Z. (2018). Exact and Robust Conformal Inference Methods for Predictive Machine Learning with Dependent Data. In *Conference On Learning Theory*, pages 732–749. PMLR. ISSN: 2640-3498.

Gibbs, I. and Candès, E. (2021). Adaptive Conformal Inference Under Distribution Shift. *arXiv:2106.00170 [stat]*. arXiv: 2106.00170.

Kath, C. and Ziel, F. (2021). Conformal prediction interval estimation and applications to day-ahead and intraday power markets. *International Journal of Forecasting*, 37(2):777–799.

Wisniewski, W., Lindsay, D., and Lindsay, S. (2020). Application of conformal prediction interval estimations to market makers' net positions. In Gammerman, A., Vovk, V., Luo, Z., Smirnov, E., and Cherubin, G., editors, *Proceedings of the Ninth Symposium on Conformal and Probabilistic Prediction and Applications*, volume 128 of *Proceedings of Machine Learning Research*, pages 285–301. PMLR.

Xu, C. and Xie, Y. (2021). Conformal prediction interval for dynamic time-series. In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 11559–11569. PMLR.

Conformal prediction and time series, what's the issue?

Time series are not exchangeable

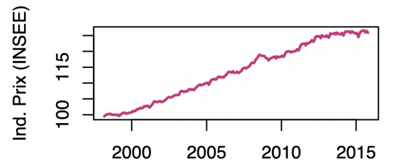


Figure 12: Trend[4]

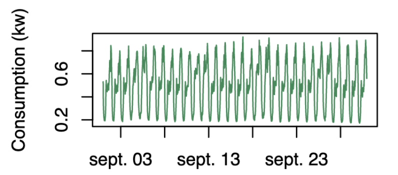


Figure 13: Seasonality[4]

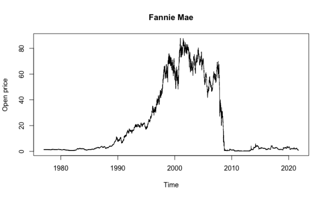


Figure 14: Shift

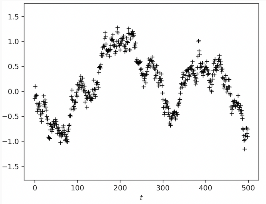


Figure 15: Time dependence

[4] Images from Yannig Goude class material.

Time dependent noise

Assume the following model:

$$Y_t = f_t(X_t) + \varepsilon_t, \text{ for } t \in \mathbb{N}^*,$$

for some function f_t, and some noise ε_t.

Time dependent noise

Assume the following model:

$$Y_t = f_t(X_t) + \varepsilon_t, \quad \text{for } t \in \mathbb{N}^*,$$

for some function f_t, and some noise ε_t.

If the noise ε_t is time dependent, the residuals will be dependent no matter what is the fitted regression function.

Time dependent noise

Assume the following model:

$$Y_t = f_t(X_t) + \varepsilon_t, \quad \text{for } t \in \mathbb{N}^*,$$

for some function f_t, and some noise ε_t.

If the noise ε_t is time dependent, the residuals will be dependent no matter what is the fitted regression function.

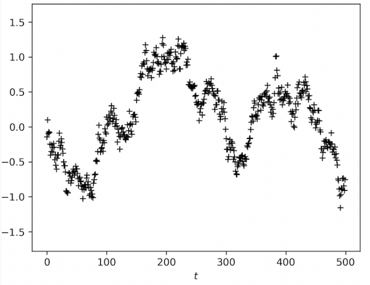


Figure 16: Auto-Regressive noise

Non-exchangeable even if the noise is exchangeable

Even if the noise is exchangeable, we can produce dependent residuals (examples available).

Endogenous and not perfectly estimated

Assume $X_t = Y_{t-1} \in \mathbb{R}$ and that

$$Y_t = aY_{t-1} + \varepsilon_t,$$

where ε_t is a white noise.

Endogenous and not perfectly estimated

Assume $X_t = Y_{t-1} \in \mathbb{R}$ and that

$$Y_t = aY_{t-1} + \varepsilon_t,$$

where ε_t is a white noise.

Assume that the fitted model is $\hat{f}_t(x) = \hat{a}x$, with $\hat{a} \neq a$.

Endogenous and not perfectly estimated

Assume $X_t = Y_{t-1} \in \mathbb{R}$ and that

$$Y_t = aY_{t-1} + \varepsilon_t,$$

where ε_t is a white noise.

Assume that the fitted model is $\hat{f}_t(x) = \hat{a}x$, with $\hat{a} \neq a$.

Then, for any t, we have that:

$$\hat{\varepsilon}_t = Y_t - \hat{Y}_t = (a - \hat{a})\, Y_{t-1} + \varepsilon_t$$

$$\hat{\varepsilon}_t = a\hat{\varepsilon}_{t-1} + \xi_t$$

with $\xi_t = \varepsilon_t - \hat{a}\varepsilon_{t-1}$.

Endogenous and not perfectly estimated

Assume $X_t = Y_{t-1} \in \mathbb{R}$ and that

$$Y_t = aY_{t-1} + \varepsilon_t,$$

where ε_t is a white noise.

Assume that the fitted model is $\hat{f}_t(x) = \hat{a}x$, with $\hat{a} \neq a$.

Then, for any t, we have that:

$$\hat{\varepsilon}_t = Y_t - \hat{Y}_t = (a - \hat{a})\, Y_{t-1} + \varepsilon_t$$

$$\hat{\varepsilon}_t = a\hat{\varepsilon}_{t-1} + \xi_t$$

with $\xi_t = \varepsilon_t - \hat{a}\varepsilon_{t-1}$.

$\hat{\varepsilon}_t$ is an ARMA process of parameters $\varphi = a$ and $\theta = -\hat{a}$.

Endogenous and not perfectly estimated

Assume $X_t = Y_{t-1} \in \mathbb{R}$ and that

$$Y_t = aY_{t-1} + \varepsilon_t,$$

where ε_t is a white noise.

Assume that the fitted model is $\hat{f}_t(x) = \hat{a}x$, with $\hat{a} \neq a$.

Then, for any t, we have that:

$$\hat{\varepsilon}_t = Y_t - \hat{Y}_t = (a - \hat{a})\, Y_{t-1} + \varepsilon_t$$

$$\hat{\varepsilon}_t = a\hat{\varepsilon}_{t-1} + \xi_t$$

with $\xi_t = \varepsilon_t - \hat{a}\varepsilon_{t-1}$.

$\hat{\varepsilon}_t$ is an ARMA process of parameters $\varphi = a$ and $\theta = -\hat{a}$.

Thus, we have generated dependent residuals (ARMA residuals) even if the underlying model only had white noise.

Exogenous and misspecified

Assume $X_t \in \mathbb{R}^2$ and that:

$$Y_t = aX_{1,t} + bX_{2,t} + \varepsilon_t,$$

with $\varepsilon_t \underset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$, $X_{2,t+1} = \varphi X_{2,t} + \xi_t,\ \ \xi_t \underset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ and $X_{1,t}$ can be any random variable.

Exogenous and misspecified

Assume $X_t \in \mathbb{R}^2$ and that:

$$Y_t = aX_{1,t} + bX_{2,t} + \varepsilon_t,$$

with $\varepsilon_t \underset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$, $X_{2,t+1} = \varphi X_{2,t} + \xi_t,\ \ \xi_t \underset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ and $X_{1,t}$ can be any random variable.

Assume that we misspecify the model such that the fitted model is $\hat{f}_t(x) = ax_1$.

Exogenous and misspecified

Assume $X_t \in \mathbb{R}^2$ and that:

$$Y_t = aX_{1,t} + bX_{2,t} + \varepsilon_t,$$

with $\varepsilon_t \underset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$, $X_{2,t+1} = \varphi X_{2,t} + \xi_t,\ \ \xi_t \underset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ and $X_{1,t}$ can be any random variable.

Assume that we misspecify the model such that the fitted model is $\hat{f}_t(x) = ax_1$.

Then, for any t, we have that

$$\hat{\varepsilon}_t = Y_t - \hat{Y}_t = bX_{2,t} + \varepsilon_t.$$

Exogenous and misspecified

Assume $X_t \in \mathbb{R}^2$ and that:

$$Y_t = aX_{1,t} + bX_{2,t} + \varepsilon_t,$$

with $\varepsilon_t \underset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$, $X_{2,t+1} = \varphi X_{2,t} + \xi_t$, $\xi_t \underset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ and $X_{1,t}$ can be any random variable.

Assume that we misspecify the model such that the fitted model is $\hat{f}_t(x) = ax_1$.

Then, for any t, we have that

$$\hat{\varepsilon}_t = Y_t - \hat{Y}_t = bX_{2,t} + \varepsilon_t.$$

Thus, we have generated dependent residuals (auto-regressive residuals) even if the underlying model only had i.i.d. Gaussian noise.

Summary of the methods

Theoretical summary of the methods

	Scores distribution		
Methods	Exchangeable	Strongly mixing	No assumption
OSCP	✓	✓[5]	✗
EnbPI	✗	✗	✗
ACI	✓	✓	✓

Table 4: Methods validity with respect to the conformity scores distribution. Green marks indicates finite-sample validity, orange long-term validity and red no theoretical validity.

[5] Chernozhukov et al. (2018)

Details on the simulation set up

Data generation

$$Y_t = 10\sin\left(\pi X_{t,1} X_{t,2}\right) + 20\left(X_{t,3} - 0.5\right)^2 + 10 X_{t,4} + 5 X_{t,5} + \varepsilon_t$$

where the X_t are multivariate uniformly distributed on $[0,1]$ and ε_t are generated from an ARMA(1,1) process.

Data generation

$$Y_t = 10\sin\left(\pi X_{t,1} X_{t,2}\right) + 20\left(X_{t,3} - 0.5\right)^2 + 10X_{t,4} + 5X_{t,5} + \varepsilon_t$$

where the X_t are multivariate uniformly distributed on $[0, 1]$ and ε_t are generated from an ARMA(1,1) process.

$\Rightarrow$ dependence structure in the noise in order to:

- control the strength of the scores dependence,
- evaluate the impact of this temporal dependence structure of the results.

Auto-Regressive Moving Average

Definition (ARMA(1,1) process)

We say that ε_t is an ARMA(1,1) process if for any t:

$$\varepsilon_{t+1} = \varphi\varepsilon_t + \xi_{t+1} + \theta\xi_t,$$

with:

- $\theta + \varphi \neq 0$, $|\varphi| < 1$ and $|\theta| < 1$;
- ξ_t is a white noise of variance σ^2, called the **innovation**.

Auto-Regressive Moving Average

Definition (ARMA(1,1) process)

We say that ε_t is an ARMA(1,1) process if for any t:

$$\varepsilon_{t+1} = \varphi\varepsilon_t + \xi_{t+1} + \theta\xi_t,$$

with:

- $\theta + \varphi \neq 0$, $|\varphi| < 1$ and $|\theta| < 1$;
- ξ_t is a white noise of variance σ^2, called the **innovation**.

- The higher φ and θ, the stronger the dependence.
- The asymptotic variance of this process is:
$$\mathrm{Var}(\varepsilon_t) = \sigma^2 \frac{1 - 2\varphi\theta + \theta^2}{1 - \varphi^2}.$$
- If $\theta = 0$, only the auto-regressive part, it is an AR(1).
- If $\varphi = 0$, only the moving-average part, it is an MA(1).

Simulation settings

- φ and θ range in $[0.1, 0.8, 0.9, 0.95, 0.99]$.
- We fix σ so as to keep the variance $\mathrm{Var}(\varepsilon_t)$ constant to 1 or 10.
- We use random forest as regressor.

Simulation settings

- φ and θ range in $[0.1, 0.8, 0.9, 0.95, 0.99]$.
- We fix σ so as to keep the variance $\mathrm{Var}(\varepsilon_t)$ constant to 1 or 10.
- We use random forest as regressor.

For each setting:

- 300 points, the last 100 kept for prediction and evaluation,
- 500 repetitions,

$\Rightarrow$ in total, $100 \times 500 = 50000$ predictions are evaluated.

Simulation settings

- φ and θ range in $[0.1, 0.8, 0.9, 0.95, 0.99]$.
- We fix σ so as to keep the variance $\mathrm{Var}(\varepsilon_t)$ constant to 1 or 10.
- We use random forest as regressor.

For each setting:

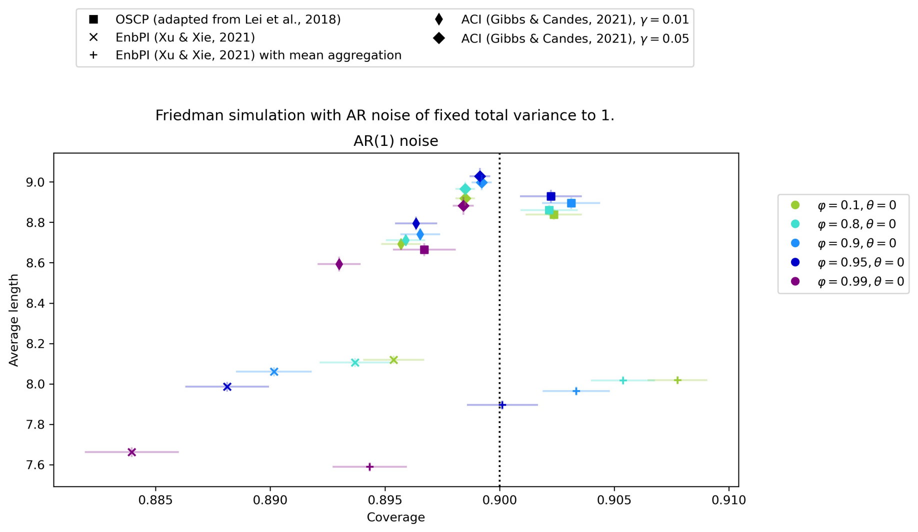
- 300 points, the last 100 kept for prediction and evaluation,
- 500 repetitions,

$\Rightarrow$ in total, $100 \times 500 = 50000$ predictions are evaluated.

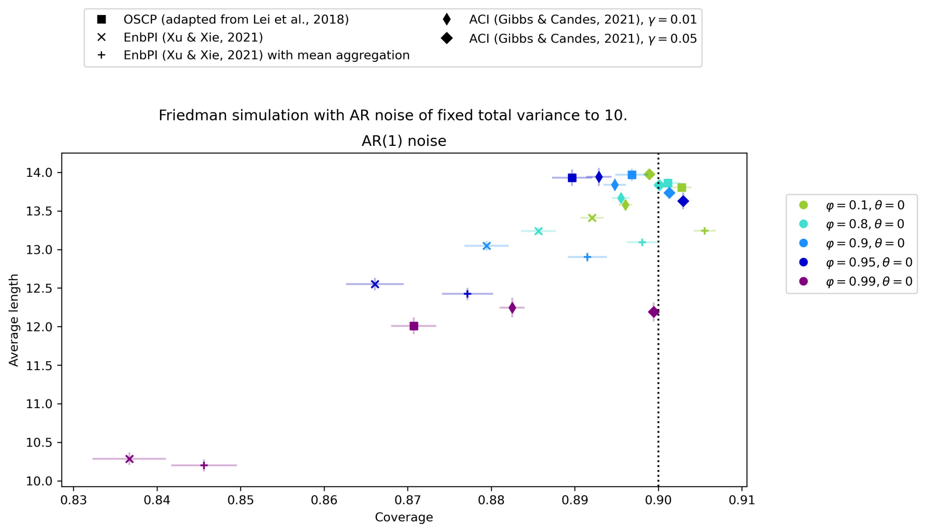
We present the results in the ARMA(1,1) case, but we also have them for AR(1) and MA(1) processes.

Additional results

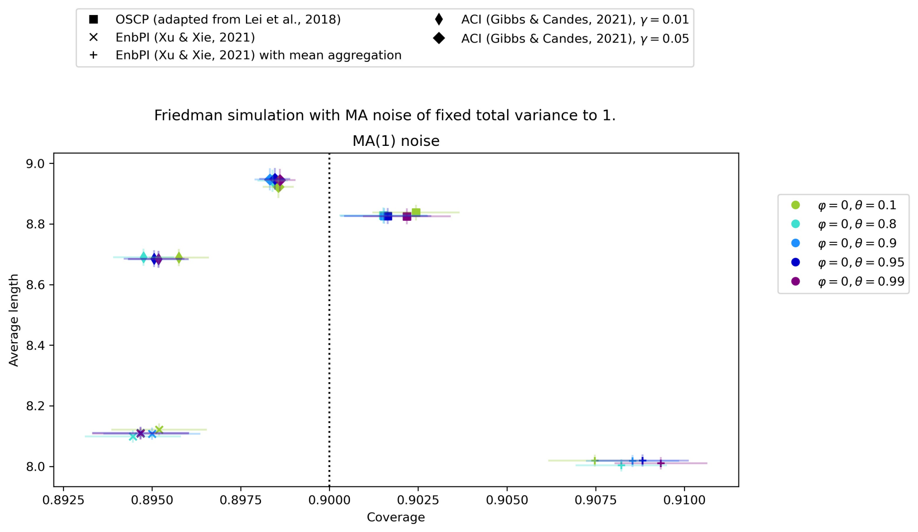
Results: impact of the temporal dependence, AR(1), variance 1



Results: impact of the temporal dependence, AR(1), variance 10



Results: impact of the temporal dependence, MA(1), variance 1



Results: impact of the temporal dependence, MA(1), variance 10