Adaptive Conformal Predictions for Time Series

An application to forecasting French electricity Spot prices

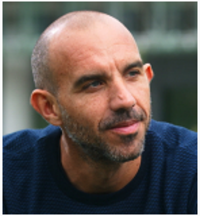
Margaux Zaffran

7th Mathematical Statistics Day – Informal Conference on Conformal Inference

Olivier Féron

EDF R&D

FiME



Yannig Goude

EDF R&D

LMO

Julie Josse

PreMeDICaL

INRIA

Aymeric Dieuleveut

École Polytechnique

Electricity Spot prices

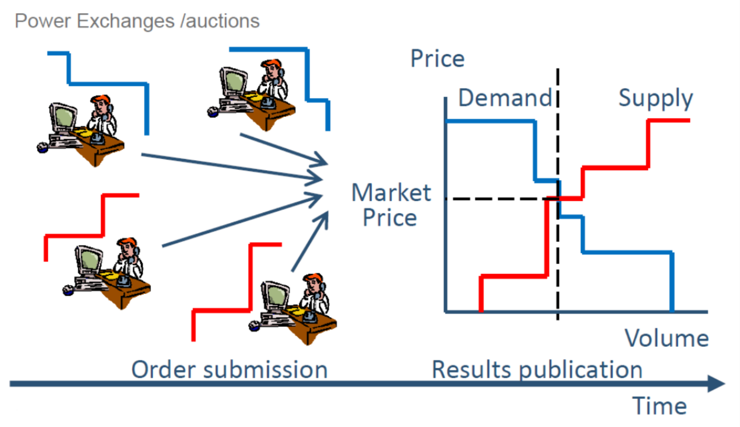


Figure 1: Drawing of spot auctions mechanism

French Electricity Spot prices data set: visualisation

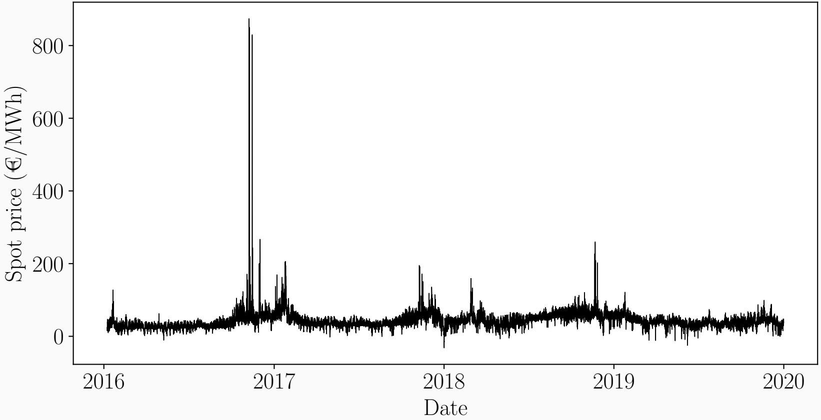


Figure 2: Representation of the French electricity spot prices, from 2016 to 2019.

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.

- $Y_t \in \mathbb{R}$
- $X_t \in \mathbb{R}^d$

Forecasting French electricity Spot prices

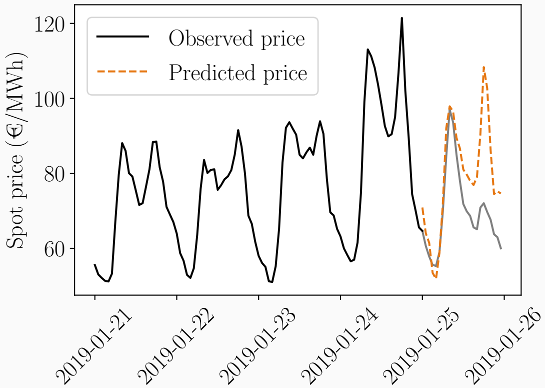


Figure 3: French electricity spot price and its prediction with random forest.

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

$\hookrightarrow$ 3 years for training

$\hookrightarrow$ 1 year to forecast

Forecasting French electricity Spot prices with confidence

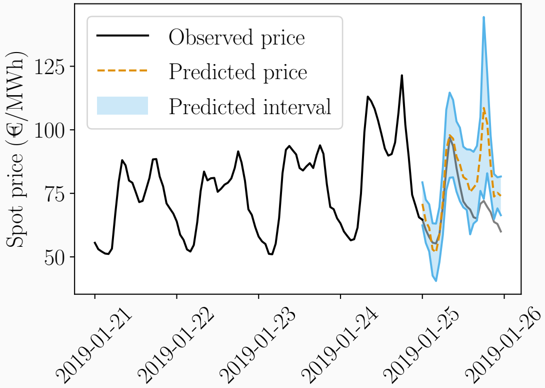


Figure 4: French electricity spot price, its prediction and its uncertainty with AgACI (proposed algorithm).

- Target coverage: 90%
- Empirical coverage: 91.68%

Generalizing beyond exchangeability in theory

Two major general theoretical results beyond exchangeability:

- Chernozhukov et al. (2018)
  $\hookrightarrow$ If the learnt model is accurate and the data noise is strongly mixing, then CP is valid asymptotically ✓
- Barber et al. (2022)
  $\hookrightarrow$ Quantifies the coverage loss depending on the strength of exchangeability violation
  $$\mathbb{P}(Y_{n+1} \in \widehat{C}_\alpha(X_{n+1})) \geq 1 - \alpha - \text{average violation of exchangeability by each calibration point}$$
  $\hookrightarrow$ proposed algorithm: reweighting (again)!
  e.g., in a temporal setting, give higher weights to more recent points.

Exchangeability does not hold in many practical applications

CP requires exchangeable data points to ensure validity

- ✗ Covariate shift, i.e. $\mathcal{L}_X$ changes but $\mathcal{L}_{Y|X}$ stays constant *(see e.g., Tibshirani et al., 2019)*
- ✗ Label shift, i.e. $\mathcal{L}_Y$ changes but $\mathcal{L}_{X|Y}$ stays constant *(see e.g., Podkopaev and Ramdas, 2021)*
- ✗ Arbitrary distribution shift *(see e.g., Cauchois et al., 2020)*
- ✗ Possibly many shifts, not only one *(main focus of this presentation)*

Online setting

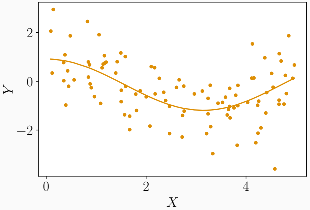
- <u>Data</u>: T_0 random variables $(X_1, Y_1), \ldots, (X_{T_0}, Y_{T_0})$ in $\mathbb{R}^d \times \mathbb{R}$
- <u>Aim</u>: predict the response values as well as predictive intervals for T_1 subsequent observations $X_{T_0+1}, \ldots, X_{T_0+T_1}$ sequentially:
  at any prediction step $t \in [\![T_0+1, T_0+T_1]\!]$, $Y_{t-T_0}, \ldots, Y_{t-1}$ have been revealed
- Build the smallest interval $\widehat{C}_\alpha^t$ such that:

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

  often simplified in:

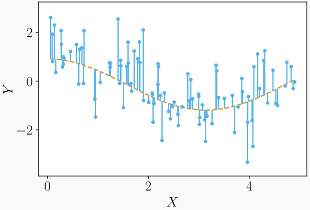
$$\frac{1}{T_1} \sum_{t=T_0+1}^{T_0+T_1} \mathbb{1}\left\{Y_t \in \widehat{C}_\alpha^t\left(X_t\right)\right\} \approx 1-\alpha.$$

1)



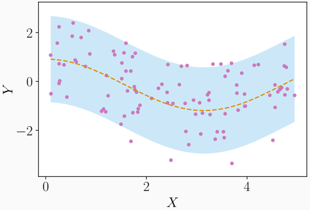
- Learn $\hat{\mu}$.

2)



- Predict with $\hat{\mu}$.
- Get the residuals $\hat{\varepsilon}_i$ and form the set of scores $\mathcal{S} = \{|\hat{\varepsilon}_i|, i \in \mathrm{Cal}\} \cup \{+\infty\}$.
- Get their $(1 - \alpha)$ empirical quantile: $Q_{1-\alpha}(\mathcal{S})$.

3)



- Predict with $\hat{\mu}$.
- Build $\hat{C}_\alpha(x)$: $[\hat{\mu}(x) \pm Q_{1-\alpha}(\mathcal{S})]$.

(Online) Time series are not exchangeable

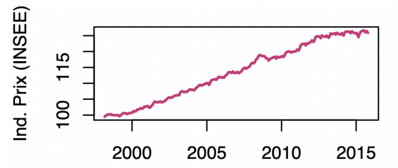


Figure 5: Trend[1]

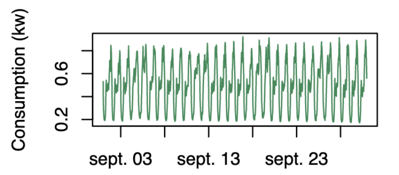


Figure 6: Seasonality[1]

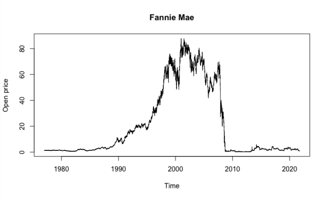


Figure 7: Shift

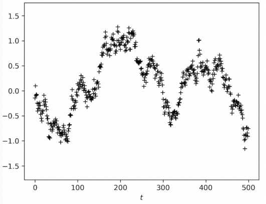


Figure 8: Time dependence

[1] Images from Yannig Goude class material.

Non-exchangeable even if the noise is exchangeable

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.

Even if the noise $(\varepsilon_t)_t$ is exchangeable, we can produce dependent residuals.

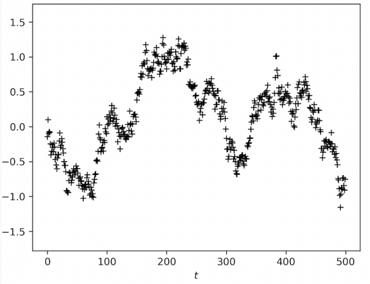


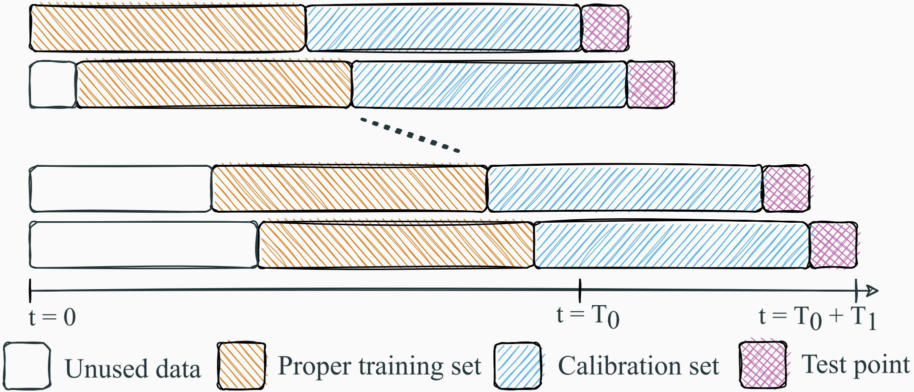
Figure 9: Auto-Regressive residuals

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)

Online sequential split conformal prediction (OSSCP)



Wisniewski et al. (2020); Kath and Ziel (2021); Zaffran et al. (2022)

$\hookrightarrow$ tested on real time series

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

Refitting the model may be insufficient ⇒ adapt the quantile level used on the calibration's scores. **(distribution shift)**

The proposed update scheme is the following:

$$\alpha_{t+1} := \alpha_t + \gamma \left(\alpha - \mathbb{1}\{ Y_t \notin \widehat{C}_{\alpha_t}(X_t) \} \right) \tag{1}$$

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

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

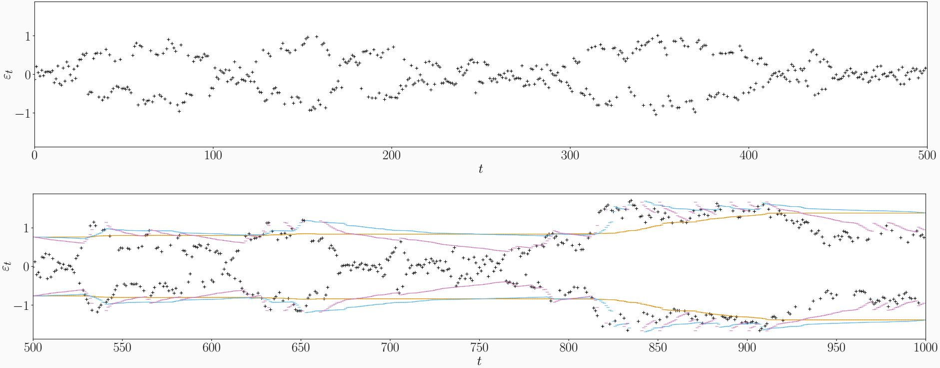


Figure 10: Visualisation of ACI with different values of γ ($\gamma = 0$, $\gamma = 0.01$, $\gamma = 0.05$)

ACI asymptotic result

Gibbs and Candès (2021) provide an asymptotic validity result for any sequence of observations.

$$\left| \frac{1}{T_1} \sum_{t=T_0+1}^{T_0+T_1} \mathbb{1}\left\{ Y_t \in \widehat{C}_{\alpha_t}(X_t) \right\} - (1-\alpha) \right| \leq \frac{2}{\gamma T_1}$$

$\Rightarrow$ favors large γ. But, the higher γ, the more frequent are the infinite intervals.

Approach

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 online context) with simple theoretical distributions
  1. exchangeable
  2. Auto-Regressive case (AR(1))
- Assume the calibration is perfect (and more), to rely on Markov Chain theory

Theoretical analysis of ACI's length: exchangeable case

Define $L(\alpha_t) = 2Q(1-\alpha_t)$ the length of the interval predicted by the adaptive algorithm at time t, and $L_0 = 2Q(1-\alpha)$ the length of the interval predicted by the non-adaptive algorithm ($\gamma = 0$).

Theorem

Assume the scores are exchangeable with quantile function Q perfectly estimated at each time, and other assumptions.

Then, for all $\gamma > 0$, $(\alpha_t)_{t>0}$ *forms a Markov Chain, that admits a stationary distribution* π_γ, *and*

$$\frac{1}{T}\sum_{t=1}^{T} L(\alpha_t) \xrightarrow[T\to+\infty]{a.s.} \mathbb{E}_{\pi_\gamma}[L] \overset{not.}{=} \mathbb{E}_{\tilde{\alpha}\sim\pi_\gamma}[L(\tilde{\alpha})].$$

Moreover, as $\gamma \to 0$,

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

Numerical analysis of ACI's length: AR(1) case

Theorem

Assume the residuals follow an AR(1) process: $\hat{\varepsilon}_{t+1} = \varphi\hat{\varepsilon}_t + \xi_{t+1}$ with $(\xi_t)_t$ i.i.d. random variables and other assumptions, we have:

$$\frac{1}{T}\sum_{t=1}^{T} L(\alpha_t) \xrightarrow[T\to+\infty]{a.s.} \mathbb{E}_{\pi_{\gamma,\varphi}}[L].$$

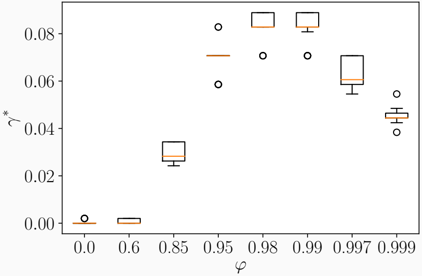
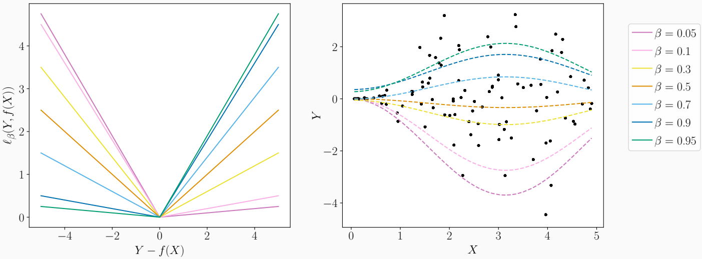


Figure 11: γ^* minimizing the average length for each φ.

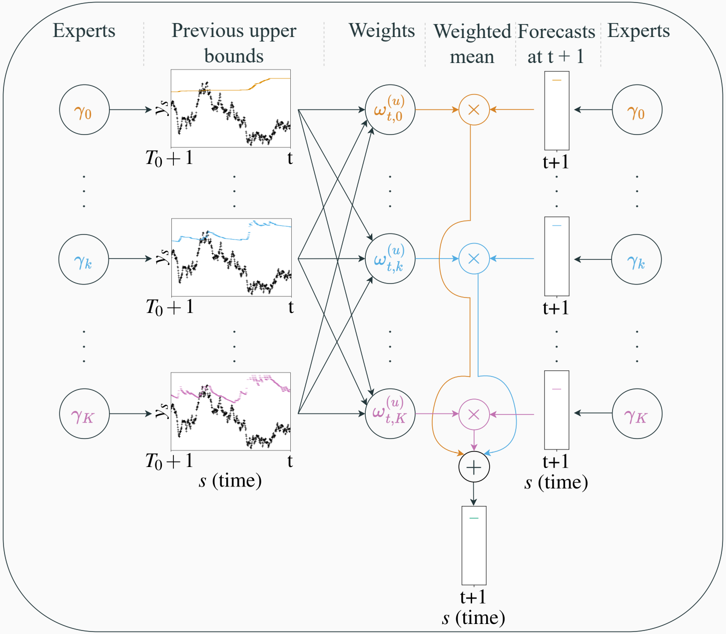
AgACI: adaptive wrapper around ACI

Online aggregation under expert advice (Cesa-Bianchi and Lugosi, 2006) computes an optimal weighted mean of experts.

`AgACI` performs 2 independent aggregations: one for each bound (the upper and lower ones), based on the pinball loss.



AgACI: adaptive wrapper around ACI, scheme (upper bound)



Data generation and simulation settings

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

where the $X_{t,\cdot} \sim \mathcal{U}([0,1])$ and ε_t is an ARMA(1,1) process:

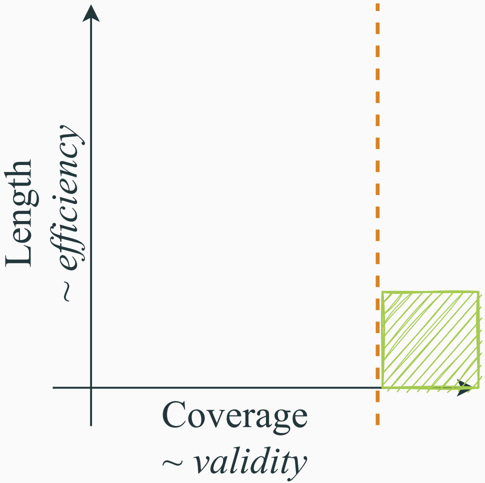
$$\varepsilon_{t+1} = \varphi\varepsilon_t + \xi_{t+1} + \theta\xi_t,$$

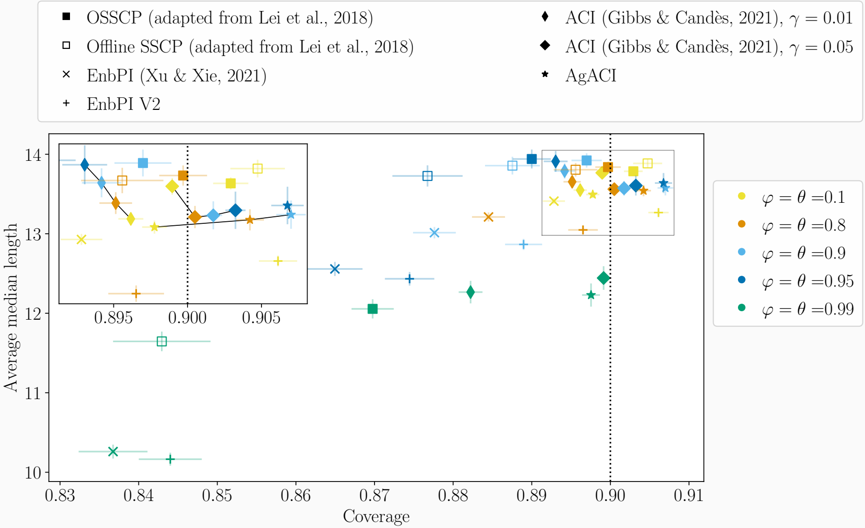
with ξ_t is a white noise of variance σ^2.

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

Visualisation of the results





OSSCP (adapted from Lei et al., 2018)
Offline SSCP (adapted from Lei et al., 2018)
EnbPI (Xu & Xie, 2021)
EnbPI V2
ACI (Gibbs & Candès, 2021), $\gamma = 0.01$
ACI (Gibbs & Candès, 2021), $\gamma = 0.05$
AgACI
$\varphi = \theta = 0.1$
$\varphi = \theta = 0.8$
$\varphi = \theta = 0.9$
$\varphi = \theta = 0.95$
$\varphi = \theta = 0.99$
Average median length
Coverage
14
13
12
11
10
0.83
0.84
0.85
0.86
0.87
0.88
0.89
0.90
0.91
0.895
0.900
0.905

Summary

1. The temporal dependence impacts the *validity*.
2. Online is significantly better than offline.
3. **OSSCP.** Achieves *valid* coverage for φ and θ smaller than 0.9, but is not robust to the increasing dependence.
4. **EnbPI.** Its *validity* strongly depends on the data distribution. When the method is *valid*, it produces the smallest intervals. EnbPI V2 method should be preferred.
5. **ACI.** Achieves *valid* coverage for every simulation settings with a well chosen γ, or for dependence such that $\varphi < 0.95$. It is robust to the strength of the dependence.
6. **AgACI.** Achieves *valid* coverage for every simulation settings, with good *efficiency*.

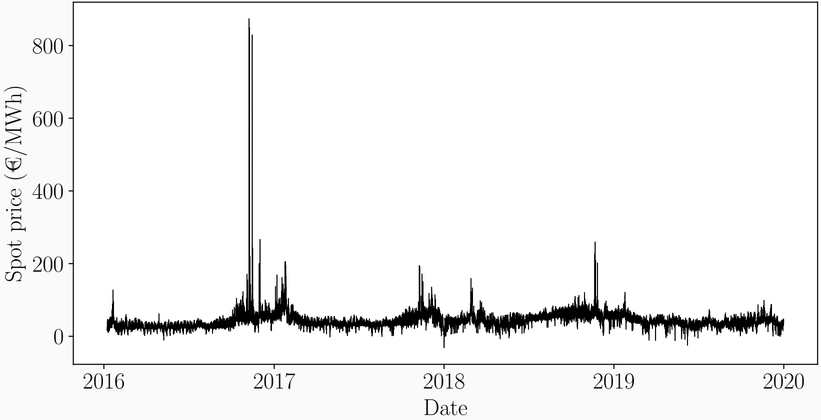


Figure 12: Representation of the French electricity spot price, from 2016 to 2019.

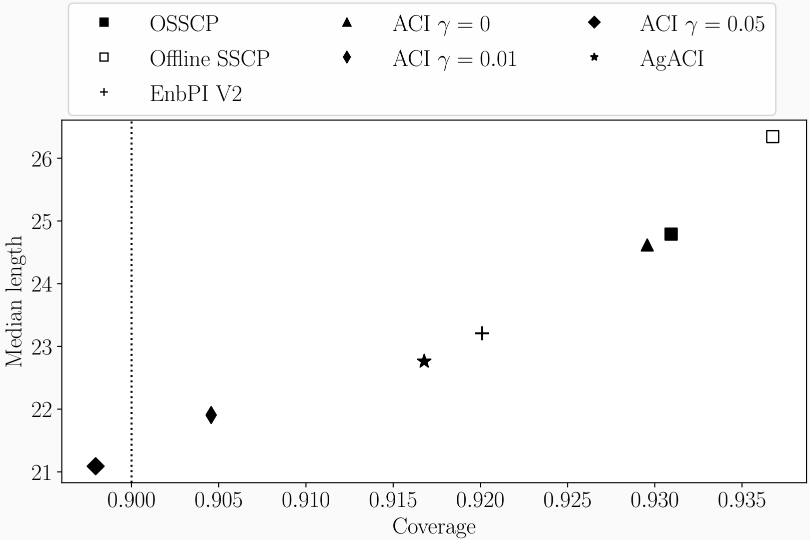
Forecasting electricity prices with confidence in 2019

- 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

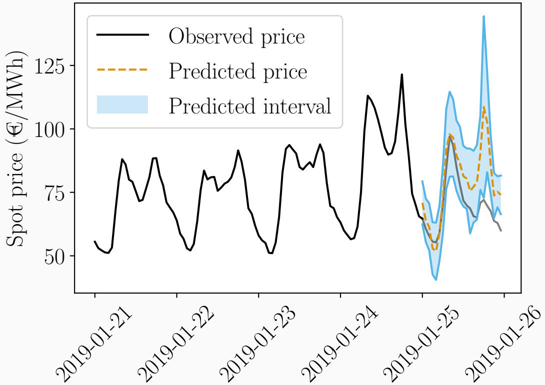


Figure 13: French electricity spot price, its prediction and its uncertainty with AgACI.

Take-home-messages

- Theoretical results on ACI's length depending on γ
- ACI useful for time series with general dependency (extensive synthetic experiments and real data)
- Empirical proposition of an adaptive choice of γ: `AgACI`

Recent developments

- Gibbs and Candès (2022) later on also proposes a method not requiring to choose γ
- Bhatnagar et al. (2023) enjoys **anytime** regret bound, by leveraging tools from the strongly adaptive regret minimization literature
- Feldman et al. (2023) extends ACI to general risk control
- Bastani et al. (2022) proposes an algorithm achieving stronger coverage guarantees (conditional on specified overlapping subsets, and threshold calibrated) without hold-out set
- Angelopoulos et al. (2023) combines CP ideas with control theory ones, to adaptively improve the predictive intervals depending on the errors structure

Non exhaustive references.

Questions? :)

Thanks for listening and feel free to reach out!

Paper $\longrightarrow$

Code $\longrightarrow$

Summary $\longrightarrow$

Angelopoulos, A. N., Candès, E. J., and Tibshirani, R. J. (2023). Conformal pid control for time series prediction. arXiv: 2307.16895.

Barber, R. F., Candès, E. J., Ramdas, A., and Tibshirani, R. J. (2022). Conformal prediction beyond exchangeability. To appear in *Annals of Statistics (2023)*.

Bastani, O., Gupta, V., Jung, C., Noarov, G., Ramalingam, R., and Roth, A. (2022). Practical adversarial multivalid conformal prediction. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc.

Bhatnagar, A., Wang, H., Xiong, C., and Bai, Y. (2023). Improved online conformal prediction via strongly adaptive online learning. In *Proceedings of the 40th International Conference on Machine Learning*. PMLR.

Cauchois, M., Gupta, S., Ali, A., and Duchi, J. C. (2020). Robust Validation: Confident Predictions Even When Distributions Shift. arXiv: 2008.04267.

Cesa-Bianchi, N. and Lugosi, G. (2006). *Prediction, learning, and games*. Cambridge University Press.

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*. PMLR.

Feldman, S., Ringel, L., Bates, S., and Romano, Y. (2023). Achieving risk control in online learning settings. *Transactions on Machine Learning Research (TMLR)*.

Gibbs, I. and Candès, E. (2021). Adaptive conformal inference under distribution shift. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc.

Gibbs, I. and Candès, E. (2022). Conformal inference for online prediction with arbitrary distribution shifts. arXiv: 2208.08401.

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).

Podkopaev, A. and Ramdas, A. (2021). Distribution-free uncertainty quantification for classification under label shift. In *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*. PMLR.

Tibshirani, R. J., Barber, R. F., Candes, E., and Ramdas, A. (2019). Conformal Prediction Under Covariate Shift. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc.

Vovk, V., Gammerman, A., and Shafer, G. (2005). *Algorithmic Learning in a Random World*. Springer US.

Wisniewski, W., Lindsay, D., and Lindsay, S. (2020). Application of conformal prediction interval estimations to market makers' net positions. In *Proceedings of the Ninth Symposium on Conformal and Probabilistic Prediction and Applications*, volume 128. 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.

Zaffran, M., Féron, O., Goude, Y., Josse, J., and Dieuleveut, A. (2022). Adaptive conformal predictions for time series. In *Proceedings of the 39th International Conference on Machine Learning*. PMLR.

Examples of non-exchangeable scores with exchangeable 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 $\hat{a} \neq a$.

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.

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.

Analysis of ACI's efficiency depending on γ

Numerical analysis of ACI's length: AR(1) case

Assume the residuals follow an AR(1) process: $\hat{\varepsilon}_{t+1} = \varphi\hat{\varepsilon}_t + \xi_{t+1}$ with $(\xi_t)_t$ i.i.d. random variables and other assumptions, we have:

$$\frac{1}{T}\sum_{t=1}^{T} L(\alpha_t) \xrightarrow[T\to+\infty]{a.s.} \mathbb{E}_{\pi_{\gamma,\varphi}}[L].$$

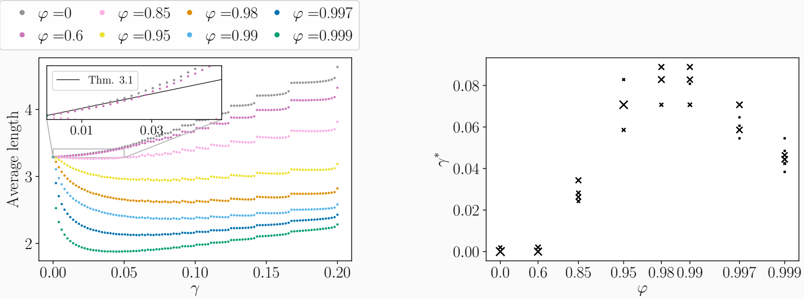


Figure 14: Left: evolution of the mean length depending on γ for various φ. Right: γ^* minimizing the average length for each φ.

EnbPI

EnbPI, Xu and Xie (2021)

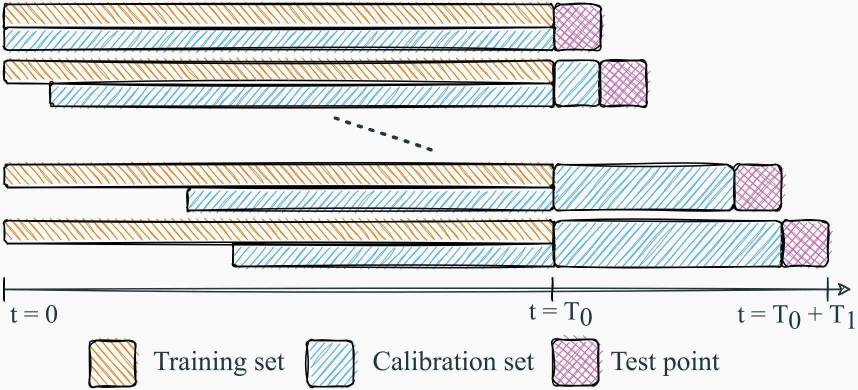


Figure 15: Diagram describing the EnbPI algorithm.

↪ tested on other real time series

↪ compared to offline methods

EnbPI (ICML, Xu and Xie, 2021) aggregates with 2 different functions.

⇒ We propose `EnbPI V2` with the same aggregation function all along (similar to

Details on the simulation set up

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.

$\Rightarrow$ 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**.

- 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.

For each setting:

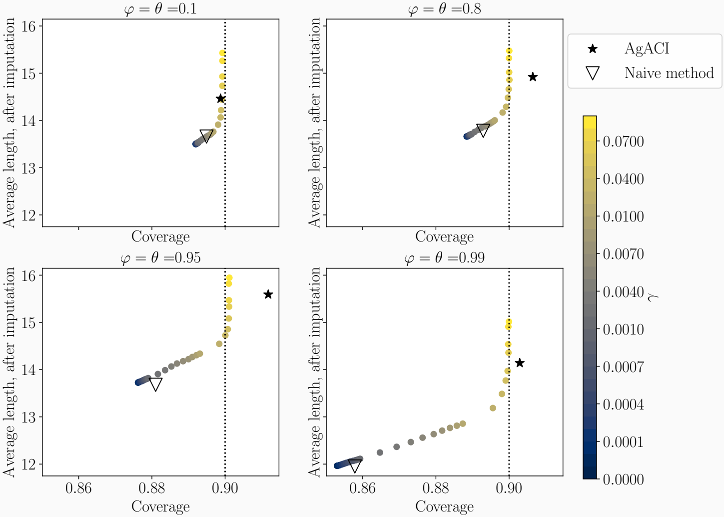
- 300 points, the last 100 kept for prediction and evaluation,
- 500 repetitions,

$\Rightarrow$ 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.

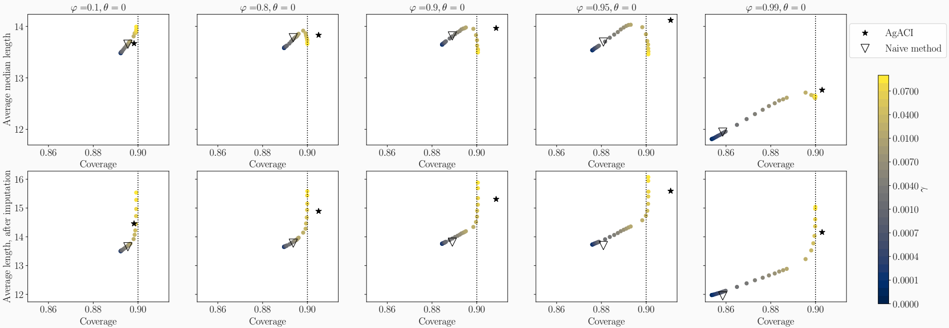
Additional results on the synthetic data sets

Empirical evaluation of ACI sensitivity to γ and adaptive choice

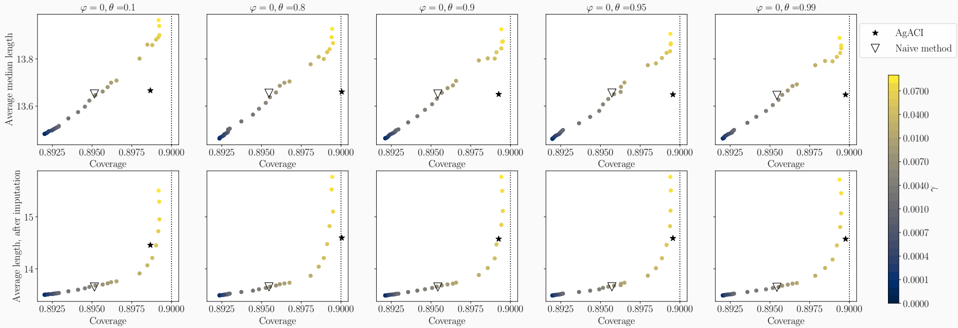


$\Rightarrow$ The more the dependence, the more sensitive to γ is ACI. Naive method ($\triangledown$): smallest among valid ones in the past $\Rightarrow$ accumulates error of the different ACI's versions. AgACI ($\bigstar$): encouraging preliminary results.

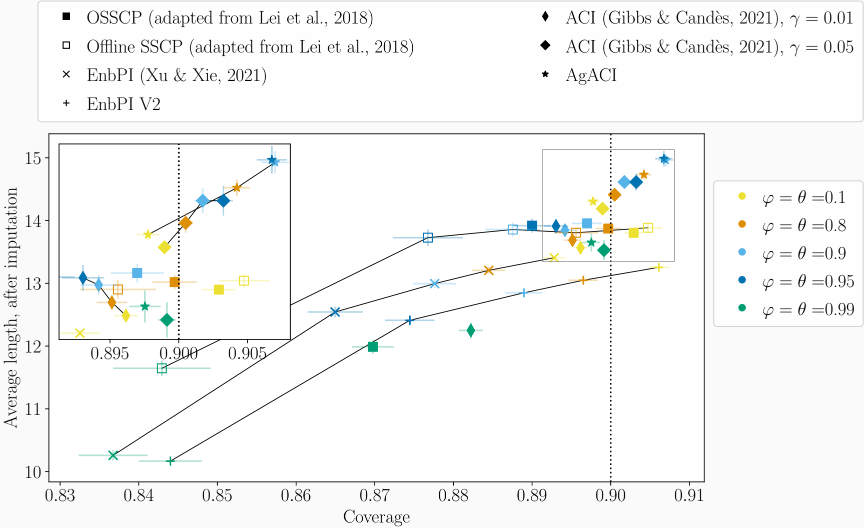
Empirical evaluation of ACI sensitivity to γ and adaptive choice, AR(1)



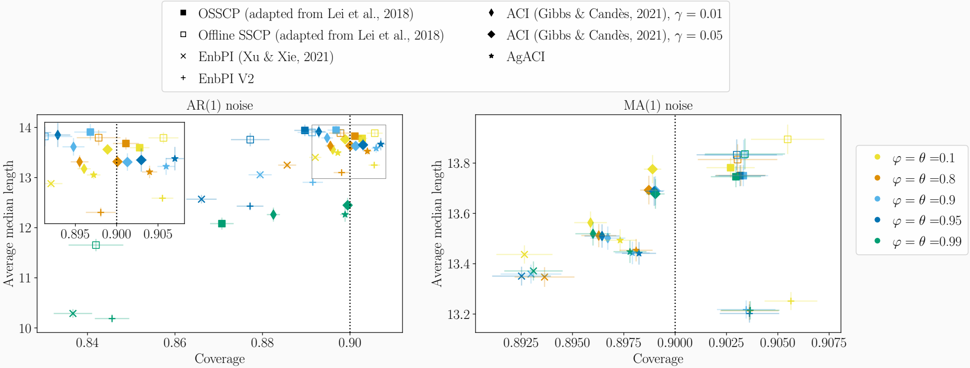
Empirical evaluation of ACI sensitivity to γ and adaptive choice, MA(1)



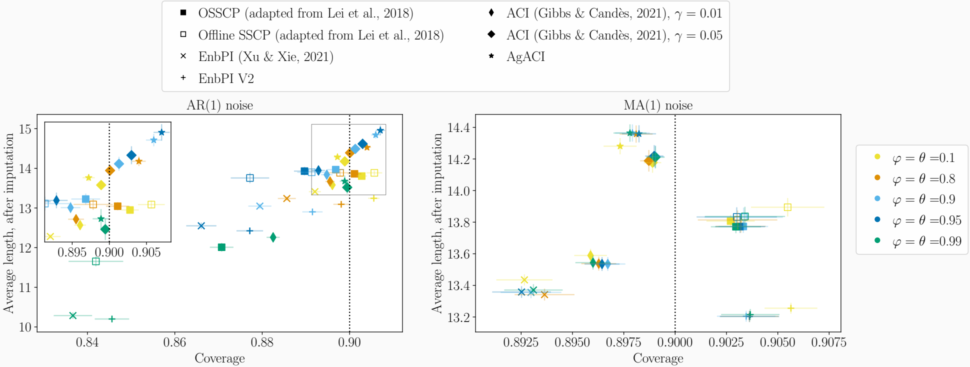
Results: impact of the temporal dependence, ARMA(1), variance 10, average length after imputation



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



Results: impact of the temporal dependence, AR(1) and MA(1), variance 10, average length after imputation



Additional results on the French electricity spot prices

Forecasting French electricity Spot prices with confidence: results

- Target coverage: 90%
- Empirical coverage: 91.68%
- Median length: 22.76€/MWh

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

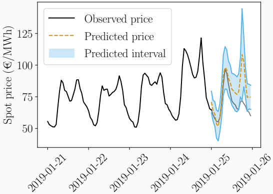


Figure 16: OSSCP

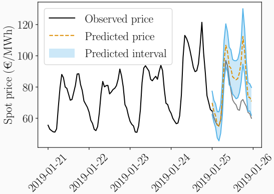


Figure 17: EnbPI V2

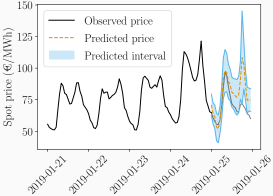


Figure 18: ACI with $\gamma = 0.01$

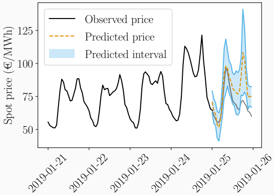


Figure 19: ACI with $\gamma = 0.05$