Physical explanations of AI time series forecasting using BAPC

Alfredo López[1] and Florian Sobieczky[1]

Software Competence Center Hagenberg, Hagenberg, Austria
alfredo.lopez@scch.at

Abstract. We introduce a local surrogate approach for explainable time-series forecasting. An initially non-interpretable predictive model to improve the forecast of a classical time-series 'base model' is used. 'Explainability' of the correction is provided by fitting the base model again to the data from which the error prediction is removed (subtracted), yielding a difference in the model parameters which can be interpreted. We provide an illustrative examples to demonstrate the potential of the method to discover and explain underlying patterns in the data.

Keywords: XAI · surrogate modeling · time-series forecasting · physics informing ML.

1 Introduction

Explainable AI (XAI) has seen a growing research interest as the demand for accountability in predictive modeling for various applications has increased [4, 2, 5]. In particular, complex AI models are widely used in modern time-series applications and therefore explainability plays a fundamental role [9, 7]. Local surrogate modeling methods such as LIME [8], SHAP [14] and others [3] have delivered the possibility to interpret the action of predictive models around a specific instance, that is, in the neighborhood of a single data point in the input space. While there are many ways in which these local approaches can be used [15, 11], there is also the fundamental question of how to determine the size of the neighborhood [10].

As model agnostic local surrogate modeling is defined by the attempt to mimic any given block-box AI model locally by a different, interpretable model, uncertainties of these models may play a major role in the quality of the emergent explanations ([10], Sect. 5.1.11). From the viewpoint of defining explainability by local surrogates, it is therefore imperative to discuss *fidelity* (of the explaining model toward the one to be explained) in connection with model accuracy. Such a definition has been supplied in [13]. Similar to this work, we discuss the optimal locality of the instance to be explained.

Further approaches yielding explanations of the action of predictive models can be seen in a particular branch of physics-informed, or knowledge-guided machine learning [12]. Also, entropy-based methods are successful in explaining time series anomalies in a physical context [1]. BAPC has a characteristic way of expressing explanations: To explain a single instance in the input space, a neighborhood is selected and another fit of the interpretable base model is carried out with data modified (corrected by the AI model to be explained) on just this neighborhood. The change of the interpretable fitting parameters is then taken as the 'explanation' of the local action of the correcting AI model (see Section 2). Naturally, the choice of the neighborhood is one of the most significant aspect of the method.

The present work extends the BAPC method to time series forecasting, where an interpretable time series model is combined with a correcting high performing machine learning model. The latter is explained inside a suitably chosen time window that assumes the role of the neighborhood. Instead of intervals or ball-shaped regions in $\mathbb{R}^d$, 'closeness' is here taken to be 'close in time'. Therefore, the question about the appropriate neighborhood becomes about the size of a sliding window, accompanying the point in time at which the prediction is to be delivered and explained.

We organize this paper by defining BAPC for time series in Section Section 2, illustrating it by a physical application in Section 3 and forming our conclusion in Section 4.

2 BAPC

In this section we specialize the "Before and After prediction Parameter Comparison (BAPC)" framework introduced in [6] for the case of explainable time series forecasting. Given an input real valued time series $y = (y_t)_{t=1}^n$ of finite size $n \in \mathbb{N}$, the BAPC consists of the following three steps:

Step-1: First application of the base model. The first step of the BAPC is to apply a parametric *base model* f_θ to the time series y, leading to

$$y_t = f_\theta(x_t) + \varepsilon_t,\ t = 1, \dots, n, \tag{1}$$

where $\theta \in R^q$ is the estimated parameter, $x_t \in \mathbb{R}^p$ is a vector of explanatory variables and $\varepsilon_t := y_t - f_\theta(x_t)$, $t = 1, \dots, n$, are the residuals.
Step-2: Application of the correction model. The base model chosen in Step-1 is interpretable but lacks overall accuracy, which motivates the use of an additional correction. Therefore, in this step, we apply a *correction model* $\widehat{\varepsilon}$ to the residuals $\varepsilon_1, \dots, \varepsilon_n$ obtained in Step-1, leading to the forecast

$$\hat{y}_{n+k} := f_\theta(x_{n+k}) + \widehat{\varepsilon}(x_{n+k}),\ k = 1, \dots, s, \tag{2}$$

where $s \in \mathbb{N}$ is a *forecast-horizon* value. Up to this point, we have combined a base model and a correction model to generate the forecast (2). The explainability is brought in the following step.
Step-3: Second application of the base model. In this step, we take a suitable *correction-window* value $r \in \mathbb{N}_0$ and compute the modified time series

$y' := (y'_t)_{t=1}^n$ defined as $y'_t = y_t$ if $1 \leq t \leq n - r$ and $y'_t = y_t - \widehat{\varepsilon}(x_t)$ if $n - r < t \leq n$. We then fit again the base model f_θ to y' leading to estimated parameter $\theta_r \in \mathrm{I\!R}^q$, the *explanation*

$$\varDelta\theta_r := \theta - \theta_r \in \mathrm{I\!R}^q \tag{3}$$

and the *surrogate model* $f_r := f_\theta + \varDelta f_r$, where $\varDelta f_r := f_\theta - f_{\theta_r}$. Strictly speaking, one would have to call $\varDelta f_r$ the surrogate because it surrogates the correction $\widehat{\varepsilon}$. However, f_r surrogates the complete model $f_\theta + \widehat{\varepsilon}$ which is of primary interest. Figure 1 illustrates the BAPC on a piece-wise constant time series.

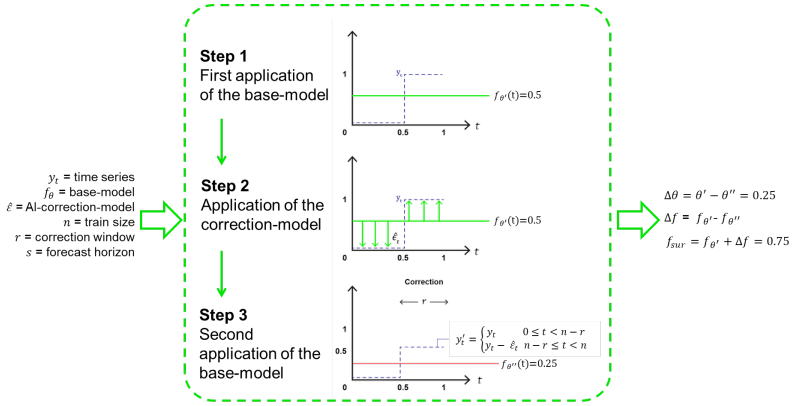
y_t = time series
f_θ = base-model
$\hat{\varepsilon}$ = AI-correction-model
n = train size
r = correction window
s = forecast horizon
Step 1
First application of the base-model
Step 2
Application of the correction-model
Step 3
Second application of the base-model
1
0
0.5
1
t
y_t
$f_{\theta'}(t)=0.5$
$\hat{\varepsilon}_t$
Correction
$\longleftarrow r \longrightarrow$
$y'_t = \begin{cases} y_t & 0 \leq t < n-r \\ y_t - \hat{\varepsilon}_t & n-r \leq t < n \end{cases}$
$f_{\theta''}(t)=0.25$
$\Delta\theta = \theta' - \theta'' = 0.25$
$\Delta f = f_{\theta'} - f_{\theta''}$
$f_{sur} = f_{\theta'} + \Delta f = 0.75$

Fig. 1. The BAPC applied to a piece-wise constant time series y having an even length n and a jump of size 1 at $(n/2)+1$. Taking as base model a constant function $f_\theta(t) = \theta$ leads to estimated parameter $\theta = 0.5$ after step-1. At step-2 we take the 1-nearest-neighborhood interpolation as correction model $\widehat{\varepsilon}$, represented by the small arrows. Taking a correction window $r = n/2$ at step-3, leads to a modified piece wise constant time series y' having a jump of size 0.5 at $(n/2)+1$. Then the parameter after correction is $\theta_r = 0.25$, leading to an explanation $\varDelta\theta_r = 0.5 - 0.25 = 0.25$ which point into the direction of $\widehat{\varepsilon}$ inside the correction window.

If observations $y_1, y_2, \ldots, y_m$ arrive consecutively over time, then it might be of interest to apply the BAPC sequentially for each point in time using the most resent data. We define the sequential BAPC (SBAPC) as the process of applying BAPC to $y_{t-n+1}, \ldots, y_t$, for $t = n, \ldots, m$ where $1 \leq n \leq m$ is a fixed training set size. The explanation at time t is given by

$$\varDelta\theta_r^t := \theta^t - \theta_r^t \in \mathrm{I\!R}^q \tag{4}$$

where θ^t and θ_r^t are the parameters obtained, respectively, after Step-1 and Step-3 of the SBAPC at time t. The surrogate model at time t is $f_r^t := f_{\theta^t} + \varDelta f_r^t$, where $\varDelta f_r^t := f_{\theta^t} - f_{\theta_r^t}$. The SBAPC is illustrated in Figure 2.

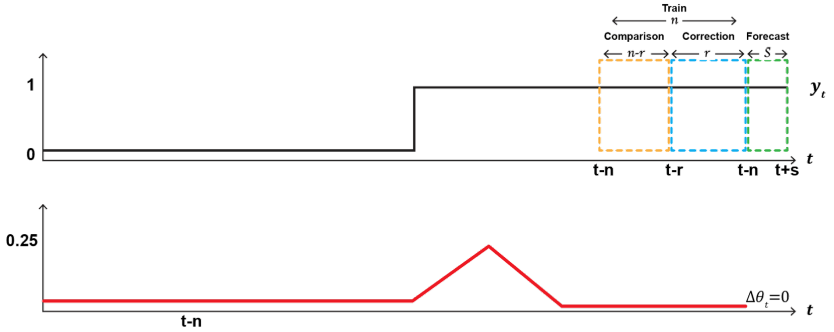

Fig. 2. The sequential-BAPC applied to a piece-wise constant step function under a similar setting than Figure 2, namely we take the constant function as base model, the 1-nearest-neighborhood interpolation as correction model and a correction window $r = n/2$ with n even. The operation on the sliding-window shown in the first row leads to the explanation $\Delta\theta_r^t$ shown in the second row.

3 Illustrative examples

In the accompanying presentation file we illustrate the proposed sequential-BAPC method by analyzing 3 synthetic time series consist of a piece-wise constant, a piece wise linear and harmonic oscillator with external force. Then we provide an example on 2023 daily SP500 stock values.

4 Conclusions

We defined an extension of the BAPC method for time series based on the notion of being able to compare the change in the model parameters of an interpretable time series model before and after the application of a complex black box correction model. In particular, we defined the concept of sequential BAPC leading to an explainable time series forecast for each point in time, by means of a time dependent explanation.

The optimal choice of the correction window (locality problem in local surrogate modeling; see Question 3 in [11]) is our primary focus of ongoing research, which we plan to propose in the form of a sequential BAPC with adaptive correction-window value. The presented illustrative example provides initial insights, by linking the explanation with the discovery and explanation of change points in the data.

It is seen that BAPC is able to deliver explanations of changes in a sequence of observations by changes of parameters associated with a law of motion if taken as a physical model for observed time series data, linking it to physics informed machine learning [12]. Furthermore, it is able to distinguish the parameters most responsible for this change from others, delivering 'feature-importance' [2] in the sense of local surrogate modeling in explainable AI.

Acknowledgments. The presented work was had been supported by the project "Secure Prescriptive Analytics", funded by the state of Upper Austria as part of the program #upperVISION2030 and by the project 'inAIco' (FFG-Project No. 862019; Bridge Young Scientist, 2020).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bukovsky, I., Kinsner, W., Homma, N.: Learning entropy as a learning-based information concept. Entropy **21**(2), 166 (2019)
2. Carvalho, D.V., Pereira, E.M., Cardoso, J.S.: Machine learning interpretability: A survey on methods and metrics. Electronics **8**(8), 832 (2019)
3. Chan-Lau, J.A., Hu, R., Ivanyna, M., Qu, R., Zhong, C.: Surrogate data models: Interpreting large-scale machine learning crisis prediction models. IMF Working Papers **2023**(041), A001 (2023)
4. Došilović, F.K., Brčić, M., Hlupić, N.: Explainable artificial intelligence: A survey. In: 2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO). pp. 0210–0215. IEEE (2018)
5. Gunning, D.: Explainable artificial intelligence (xai) darpa-baa-16-53. Defense Advanced Research Projects Agency (2016)
6. Neugebauer, S., Rippitsch, L., Sobieczky, F., Geiß, M.: Explainability of ai-predictions based on psychological profiling. Procedia Computer Science **180**, 1003–1012 (2021), proceedings of the 2nd International Conference on Industry 4.0 and Smart Manufacturing (ISM 2020)
7. Ozyegen, O., Ilic, I., Cevik, M.: Evaluation of interpretability methods for multivariate time series forecasting. Appl. Intell. **52**(5), 4727–4743 (2022)
8. Ribeiro, M.T., Singh, S., Guestrin, C.: "why should i trust you?": Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. p. 1135–1144. KDD '16, Association for Computing Machinery, New York, NY, USA (2016)
9. Rojat, T., Puget, R., Filliat, D., Del Ser, J., Gelin, R., Díaz-Rodríguez, N.: Explainable artificial intelligence (xai) on timeseries data: A survey. arXiv preprint arXiv:2104.00950 (2021)
10. Saeed, W., Omlin, C.: Explainable ai (xai): A systematic meta-survey of current challenges and future opportunities. Knowledge-Based Systems **263**, 110273 (2023)
11. Saluja, R., Malhi, A., Knapič, S., Främling, K., Cavdar, C.: Towards a rigorous evaluation of explainability for multivariate time series (2021)
12. Sel, K., Mohammadi, A., Pettigrew, R.I., Jafari, R.: Physics-informed neural networks for modeling physiological time series for cuffless blood pressure estimation. npj Digit. Medicine **6** (2023)
13. Sobieczky, F., Geiß, M.: Explainable AI by BAPC – Before and After correction Parameter Comparison. arXiv:2103.07155 (2023)
14. Štrumbelj, E., Kononenko, I.: Explaining prediction models and individual predictions with feature contributions. Knowledge and Information Systems **41**, 647–665 (2014)
15. Sundararajan, M., Najmi, A.: In: III, H.D., Singh, A. (eds.) Proceedings of the 37th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 119, pp. 9269–9278. PMLR (13–18 Jul 2020)