Stockholm University

Counterfactual Explanations for Time Series Forecasting

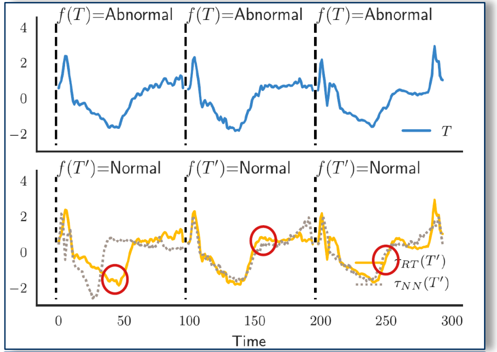
Zhendong Wang, Ioanna Miliou, Isak Samsten, and Panagiotis Papapetrou

Zhendong Wang, PhD Student

Stockholm University

Time series counterfactuals

- Counterfactual (CF) explanations:
  - Show modifications required to change a prediction from an **undesired** (e.g., unhealthy patient) to a **desired** state (e.g., healthy patient)
- Recent work in time series classification (TSC):
  - Random shapelet forest (RSF)
  - LatentCF++ (gradient-based)
  - Native Guide (instance-based)
  - And so on…



CF example on ECG Classification

How about time series forecasting?

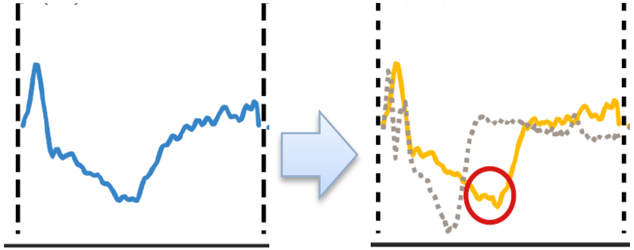
- Forecasting is highly applicable to several domains
  - E.g., sales demand forecasting, medical prognostic tasks

- Post-hoc methods like LIME, SHAP and saliency maps can be applied to provide model explainability

- However, little emphasis has been given on *actionability*, and how can a *forecasting outcome* be changed in terms of counterfactual reasoning

How to define the desired outcome?

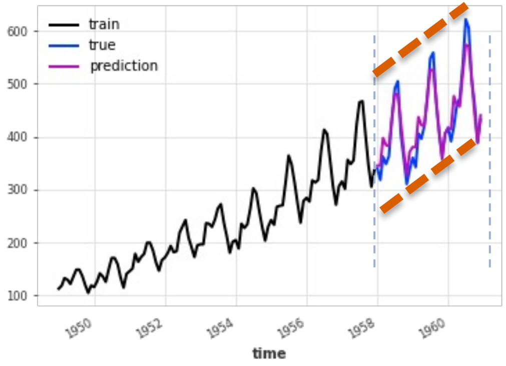
Time Series Classification



Unhealthy

Healthy

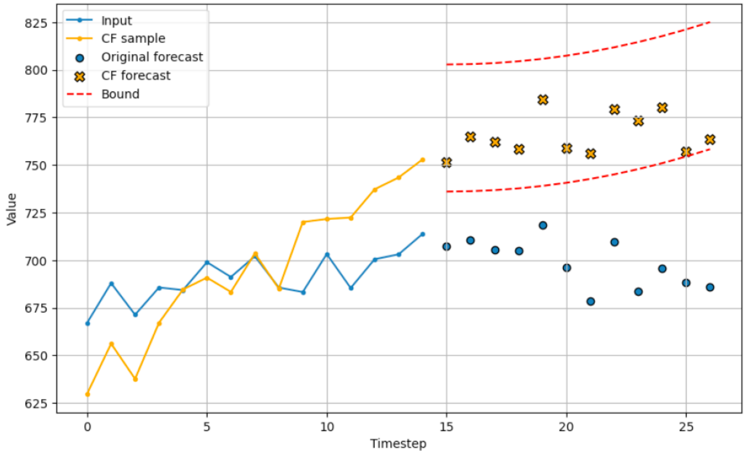
Time Series Forecasting



Define upper and lower constraints!

Example: CF for time series forecasting

- Sales forecasting of a product for the next 12 days (*blue dots*), with upper and lower constraints (*red-dotted lines*)
- Original sample illustrated in *blue*, CF sample in *yellow*



Problem formulation

- Given a black-box forecaster:

$$\hat{f}(< x_{n-d+1}, \dots, x_{n-1}, x_n >) =< \hat{x}_{n+1}, \hat{x}_{n+2}, \dots, \hat{x}_{n+T} >$$

Objective: to modify

$$x \to x', \; s.t. \; \hat{f}(x') = \hat{x}',$$

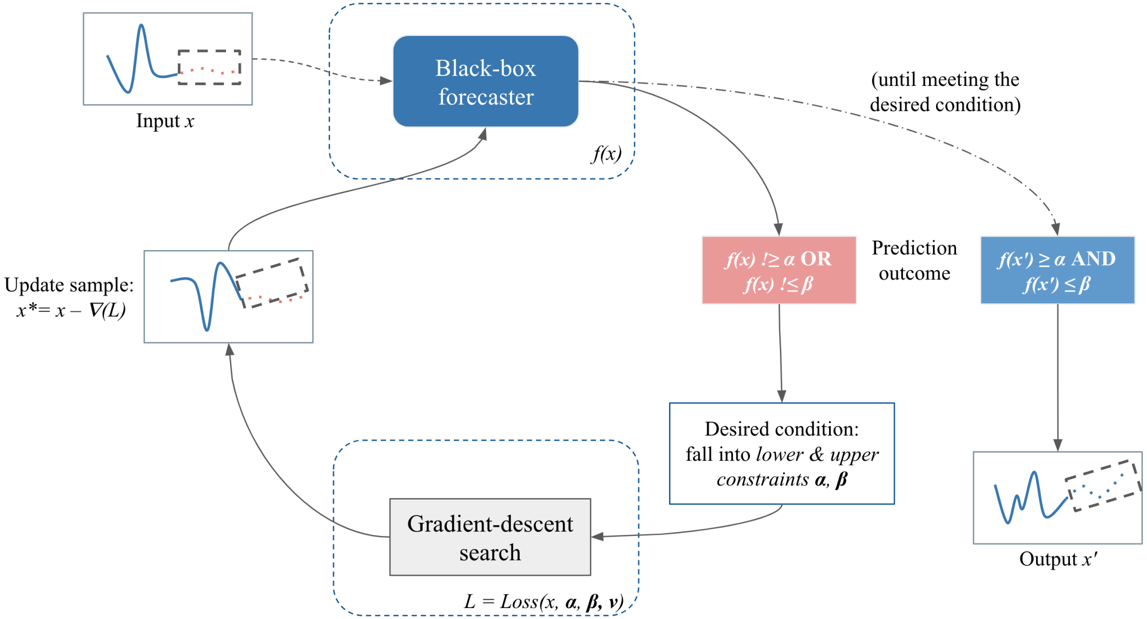
where $\hat{x}'$ falls between $\boldsymbol{\alpha} = \{\alpha_1, \dots, \alpha_T\}$ and $\boldsymbol{\beta} = \{\beta_1, \dots, \beta_T\}$.

Goal:

$$\underbrace{\boldsymbol{x}'}_{\text{Counterfactual}} = \underset{x'}{\operatorname{argmin}} \; ||\hat{f}(\boldsymbol{x}') - \underbrace{\boldsymbol{\alpha}}_{\text{Lower bound}}|| + ||\underbrace{\boldsymbol{\beta}}_{\text{Upper bound}} - \hat{f}(\boldsymbol{x}')||$$

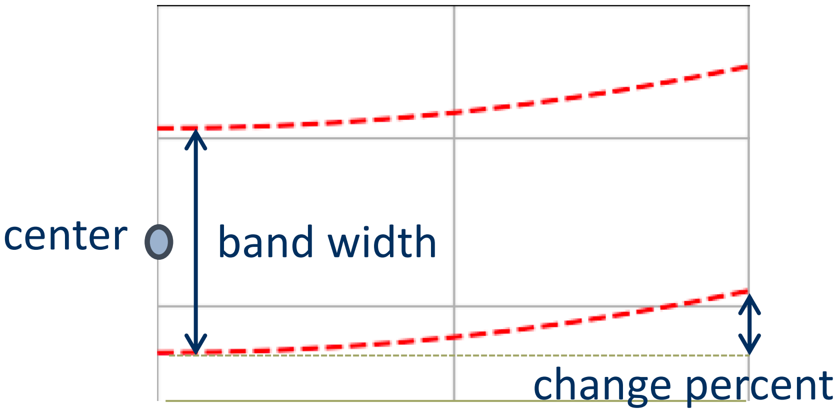
Solution: ADAM gradient optimization using a binary masking vector $\boldsymbol{v}$, containing 1's and 0's (timesteps satisfying the condition or not) in Loss

Proposed solution: ForecastCF



ForecastCF: define the desired trend

- Instantiation of polynomial trend (upper and lower bounds):



- Controlled by five hyperparameters: center $c(\cdot)$, shift s, fraction of std fr, desired change percent cp, and polynomial order $poly\_order$

Experimental set-up

- 6 real-world datasets
  - four benchmark datasets from recent competitions: *CIF2016, NN5, Tourism*, and *M4 Finance*
  - two datasets in stock marketing and healthcare: *SP500*, *MIMIC*
- 4 forecasting models
  - GRU
  - Seq2seq
  - WaveNet
  - N-Beats
- CF evaluation metrics
  - *Validity Ratio* (proportion of valid timesteps)
  - *Stepwise Validity AUC* (consecutive validity)
  - *Proximity* (Euclidean distance)
  - *Compactness* (unchanged proportion)

[Github link:]

SCAN ME

Empirical evalutaion

Stockholm University

Baseline 1:
1-nearest-
-neighbour

Baseline 2:
direct shift

Table III. Validity: *validity ratio* and *stepwise validity AUC*. We report the average of five runs and highlight the best metric in bold.

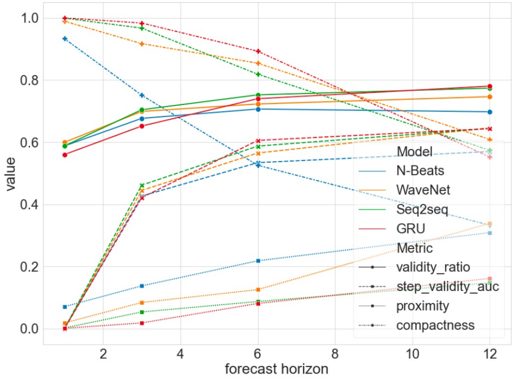
Model	CF model	CIF2016		NN5		Tourism		M4 Finance		SP500		MIMIC	
		Ratio	*S-AUC*	*Ratio*	*S-AUC*	*Ratio*	*S-AUC*	*Ratio*	*S-AUC*	*Ratio*	*S-AUC*	*Ratio*	*S-AUC*
GRU	BaseNN	0.474	0.244	0.879	0.477	0.910	0.751	0.601	0.557	0.276	0.190	0.639	0.505
	BaseShift	0.506	0.360	0.699	0.163	0.725	0.380	0.576	0.367	0.339	0.090	0.550	0.327
	ForecastCF	**0.781**	**0.650**	**1.000**	**0.980**	**0.990**	**0.922**	**0.832**	**0.655**	**0.688**	**0.338**	**0.941**	**0.800**
Seq2seq	BaseNN	0.536	0.308	0.974	0.942	0.932	0.836	0.634	0.588	0.279	0.220	0.671	0.549
	BaseShift	0.487	0.318	0.939	0.842	0.790	0.459	0.579	0.419	0.270	0.076	0.502	0.327
	ForecastCF	**0.792**	**0.667**	**0.995**	**0.973**	**0.998**	**0.953**	**0.833**	**0.686**	**0.557**	**0.320**	**0.912**	**0.760**
WaveNet	BaseNN	0.216	0.008	0.721	0.101	0.869	0.600	0.651	0.529	0.277	0.056	0.483	0.087
	BaseShift	0.332	0.043	0.552	0.024	0.622	0.117	0.530	0.149	0.224	0.026	0.382	0.043
	ForecastCF	**0.742**	**0.636**	**0.997**	**0.916**	**0.958**	**0.691**	**0.867**	**0.781**	**0.933**	**0.857**	**0.887**	**0.713**
N-Beats	BaseNN	0.531	0.291	0.885	0.375	0.924	0.705	0.650	0.552	0.325	0.149	0.634	0.398
	BaseShift	0.362	0.055	0.655	0.052	0.630	0.174	0.482	0.240	0.296	0.039	0.532	0.221
	ForecastCF	**0.699**	**0.567**	**1.000**	**0.980**	**0.984**	**0.920**	**0.884**	**0.778**	**0.879**	**0.727**	**0.928**	**0.772**

Table IV. Data manifold closeness: *proximity* and *compactness*. We report the average of five runs and highlight the best metric in bold. The † sign indicates that BaseNN and BaseShift generated the same counterfactuals across different forecasting models due to the nature of the methods.

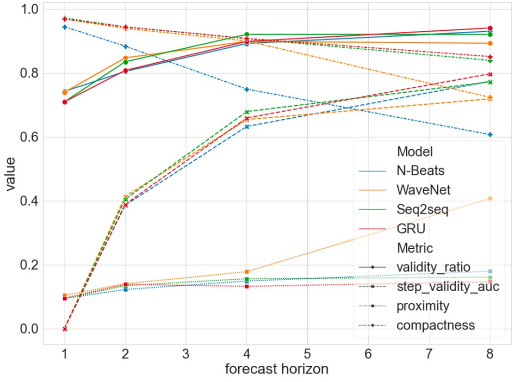
Model	CF model	CIF2016		NN5		Tourism		M4 Finance		SP500		MIMIC	
		Proxi.	*Compa.*	*Proxi.*	*Compa.*	*Proxi.*	*Compa.*	*Proxi.*	*Compa.*	*Proxi.*	*Compa.*	*Proxi.*	*Compa.*
GRU	BaseNN†	1.518	0.018	2.001	0.037	2.387	0.032	1.265	0.064	3.448	0.033	1.732	0.026
	BaseShift†	0.265	0.091	**0.384**	0.039	0.472	0.025	0.424	0.070	0.999	0.043	0.264	0.090
	ForecastCF	**0.171**	**0.534**	0.503	**0.348**	**0.323**	**0.514**	**0.172**	**0.837**	**0.660**	**0.900**	**0.153**	**0.846**
Seq2seq	BaseNN†	1.518	0.018	2.001	0.037	2.387	0.032	1.265	0.064	3.448	0.033	1.732	0.026
	BaseShift†	0.265	0.091	0.384	0.039	0.472	0.025	0.424	0.070	0.999	0.043	0.264	0.090
	ForecastCF	**0.136**	**0.593**	**0.012**	**0.979**	**0.148**	**0.653**	**0.144**	**0.844**	**0.704**	**0.875**	**0.163**	**0.839**
WaveNet	BaseNN†	1.518	0.018	2.001	0.037	2.387	0.032	1.265	0.064	3.448	0.033	1.732	0.026
	BaseShift†	**0.265**	0.091	**0.384**	0.039	**0.472**	0.025	0.424	0.070	0.999	0.043	**0.264**	0.090
	ForecastCF	0.340	**0.613**	0.912	**0.645**	0.635	**0.705**	**0.141**	**0.767**	**0.421**	**0.758**	0.408	**0.723**
N-Beats	BaseNN†	1.518	0.018	2.001	0.037	2.387	0.032	1.265	0.064	3.448	0.033	1.732	0.026
	BaseShift†	**0.265**	0.091	**0.384**	0.039	**0.472**	0.025	0.424	0.070	0.999	0.043	0.264	0.090
	ForecastCF	0.306	**0.337**	0.655	**0.080**	0.562	**0.162**	**0.131**	**0.589**	**0.512**	**0.299**	**0.183**	**0.615**

Empirical evalutaion cont.

- **Horizon test**: investigates the effectiveness of ForecastCF, the horizon increases gradually from 1 to the defined value (e.g., 12/8)
- Found a **trade-off**: higher *validity ratio* and *stepwise AUC scores*; less *proximate* and *compact* (i.e., more modifications required)



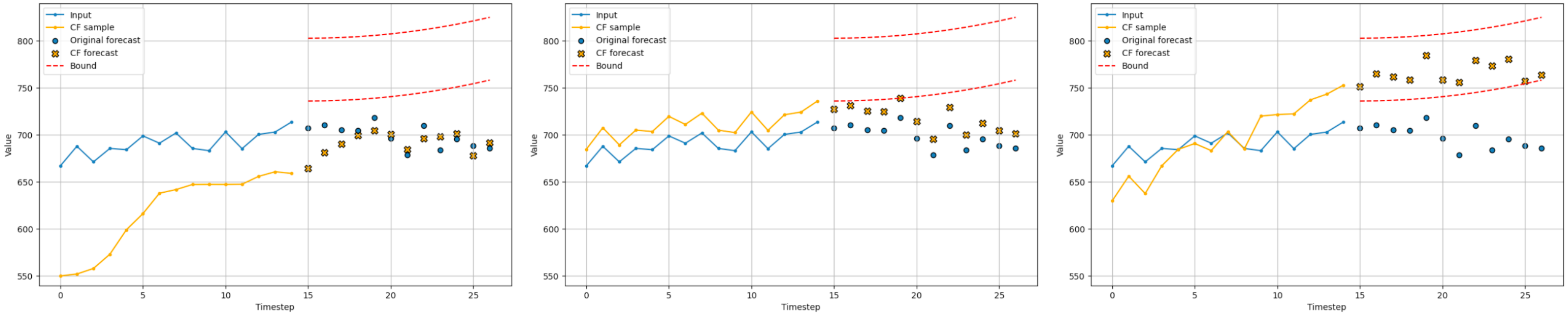
(a) CIF2016



(b) MIMIC

Qualitative analysis of examples

- In comparison with 2 baselines: *BaseNN* and *BaseShift*
- ForecastCF had the most *proximate* and *compact* counterfactual (the yellow line) compared to the other two baselines
- All 12 predicted values (yellow points) were *valid* in ForecastCF



(a) BaseNN (b) BaseShift (c) ForecastCF

Conclusions

- Summary of the paper:
  - We formulated the novel problem of counterfactual (CF) explanations for time series forecasting, and demonstrate its applicability to several application domains
  - ForecastCF: a gradient-based algorithm for generating CFs, so that the forecasted values over a time horizon satisfy a set of lower and upper bound constraints
- Future work:
  - Extending our solution into other forecasting models (e.g., traditional statistical models)
  - Investigating the multivariate forecasting setup, and incorporating exogenous variables

Github link:

SCAN ME

References

1. Wachter, S., Mittelstadt, B., Russell, C.: Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. Technical report, Social Science Research Network (2017)
2. Karlsson, I., Rebane, J., Papapetrou, P., Gionis, A.: Explainable time series tweaking via irreversible and reversible temporal transformations. In: ICDM, pp. 207-216 (2018)
3. Oreshkin, B.N., Carpov, D., Chapados, N., Bengio, Y.: N-BEATS: Neural basis expansion analysis for interpretable time series forecasting, in: ICLR. (2020)
4. Wang, Z., Samsten, I., Mochaourab, R., Papapetrou, P.: Learning Time Series Counterfactuals via Latent Space Representations, in: Discovery Science, pp. 369–384 (2021)
5. Delaney, E., Greene, D., Keane, M.T.: Instance-Based Counterfactual Explanations for Time Series Classification, in: Case-Based Reasoning Research and Development, pp. 32–47 (2021)

Many thanks for your attention!