CSI - Explainable Deep Learning for Multivariate Time Series Analytics

07/06/2023

Etienne Vareille Michele Linardi
Vassilis Christophides
ETIS, UMR8051
CY Cergy Paris Université, ENSEA

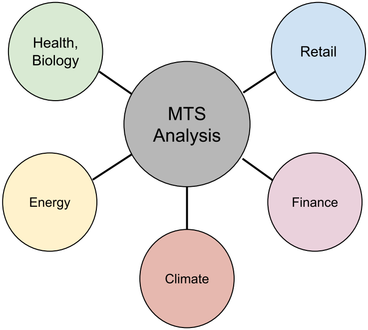
ETIS
Équipes Traitement de l'Information et Systèmes

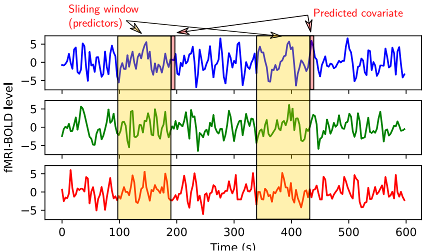
CERGY PARIS UNIVERSITÉ

Équipes Traitement
de l'Information
et Systèmes

Outline

Context - Multivariate Time Series (MTS)

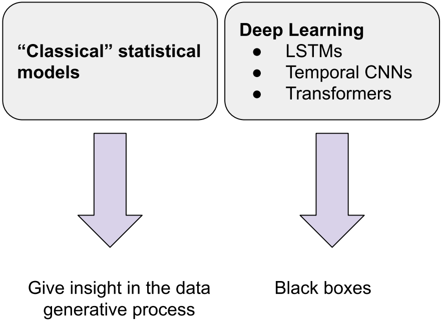




- MTS: Sequences of measured values at equally spaced instants
- Prediction task: classify/forecast next value of a variable given previous values over a window.

Équipes Traitement
de l'Information
et Systèmes

Motivation - Explaining MTS modeling



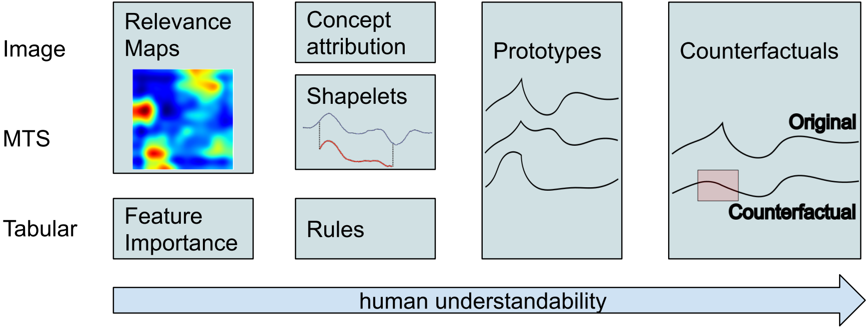
- How does a model make decision? Nauta et al. 2023

  Faithfulness | Continuity | Completeness

- How does the model relate to the data generative process? Nauta et al. 2023; Ismail et al. 2020

Équipes Traitement
de l'Information
et Systèmes

Post-Hoc Deep Learning Explicability



see Bodria et al. 2021

Figure: Most common types of explanations

Relevance attribution maps

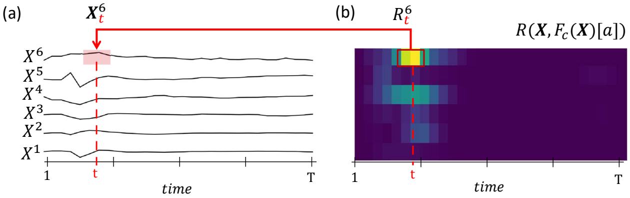


Figure: (a) Time series input, (b) Relevance Map

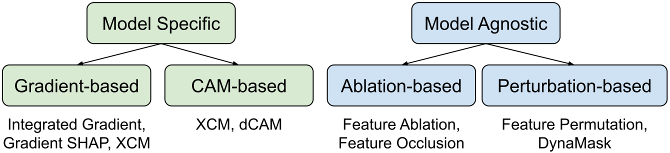


Figure: Nomenclature of relevance map methods

Causality in Multivariate Time Series

Multivariate time series as **Stochastic processes**: X_t^i is a (probabilistic) function of previous values X_{t-}.
Cause, Effect: A causes B iff (A,B) is an edge in the graph.

Consistency through time

The generative function and causal graph are identical at all timesteps.

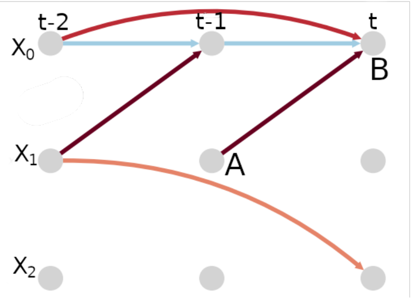


Figure: Example of a causal graph

Sufficiency of Causal Explanations

Causation and Correlation can be distinguished (direct link vs indirect link).

Sufficiency of the measured covariates

All variables that influence the process are observed.

Causal Explanations

All observed information is contained in the set of causes.
The generative process is entirely explained by the causes: hence we speak of **causal explanations**.

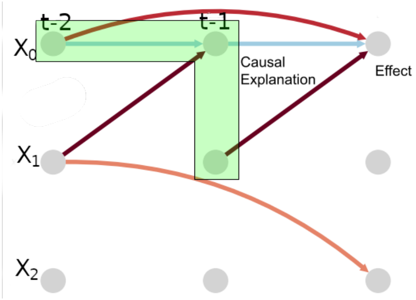


Figure: Example of a causal graph

Key questions of our empirical study

How do relevance map post-hoc xAI methods compare to causal explanations?

- I) To what extent do xAI methods explain the same way different samples in a same MTS?
- II) How much do causal explanations and relevance maps overlap?
- III) We build surrogate models restricted to the most relevant / causal features. Which has the highest predictive performance?

Experimental setting - Data, Models

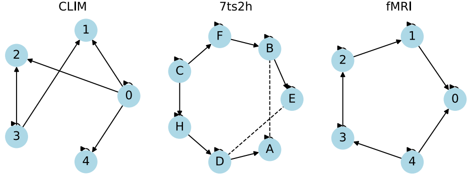


Figure: Exemple of causal graph

Dataset	Instances	Variables	Timestamps	Avg in-degree	Max lag
CLIM Runge et al. 2020	200	5	250	2.0 to 4.4	2
7ts2h Assaad, Devijver, and Gaussier 2022	10	7	4000	1.8	1
fMRI Huang and Kleinberg 2015	17	5	200 to 5000	2.0 to 2.6	1

Data properties:

- CLIM: climate, linear
- 7ts2h: nonlinear
- fMRI: health, nonlinear

Models: we explain trained models above 0.7 AUROC.

	7ts2h	CLIM	fMRI
LSTM	0.912 (52)	0.775 (115)	0.807 (49)
XCM	0.904 (45)	0.820 (45)	0.893 (15)
DCAM	0.858 (38)	0.776 (61)	0.799 (25)
Transformer	0.899 (50)	0.784 (26)	0.915 (16)

Table: AUROC and number of explained models

Experimental setting - Metrics

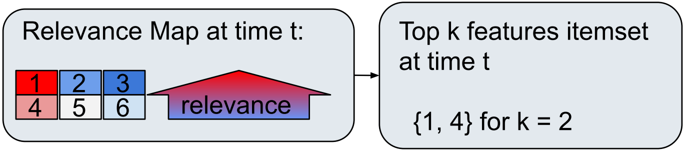


Figure: Extract itemsets

Metrics:

- Consistency is an entropy measure on the ensemble of explanations on many window.
- Precision, Recall of the itemsets compared to the causal explanation
- AUROC of surrogate models trained over masked data. The non-masked features are those belonging to 1) the causal explanation or 2) the most frequently relevant features.

Jacob et al. 2021

I - Consistency through time of explanations

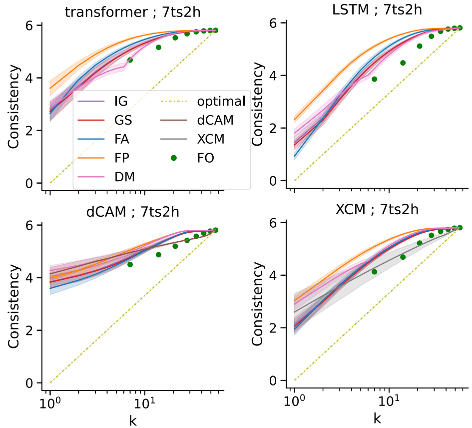


Figure: Consistency over time at different levels of k.

Observations:

- Explanations become less consistent as we include less relevant terms.
- Smaller models are more consistent, xAI methods have similar behaviours despite individual differences.

II - Precision of explanations

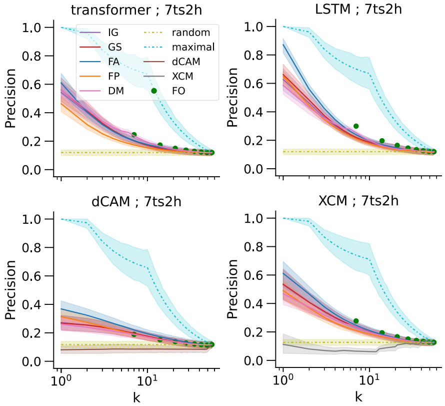


Figure: Precision over the causal graph at different levels of k.

Observations

- High precision for low *k*, gets closer to random baseline (dashed yellow) as *k* increases.
- The xAI methods have similar behaviour.
- Model type affects the precision.

Équipes Traitement
de l'Information
et Systèmes

II - Recall of explanations

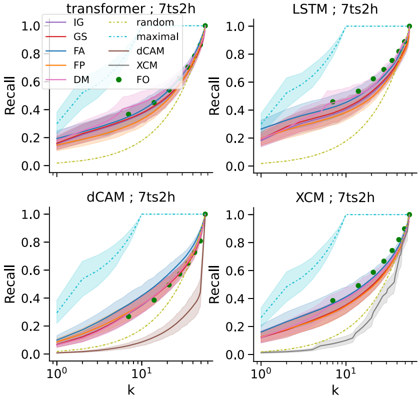


Figure: Recall over the causal graph at different levels of k.

Observations

- Same observations as those for precision.
- The most salient features ($k \leq 10$) miss at least half of the causes (down to 20% in the linear CLIM dataset)

Équipes Traitement
de l'Information
et Systèmes

III - Predictive performances of explanations

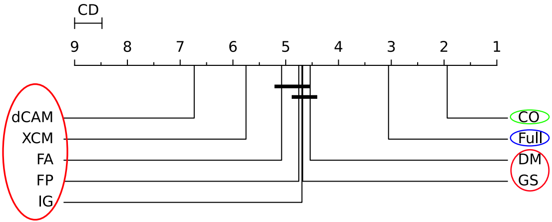


Figure: Critical Difference diagram of the AUROC of surrogate models

- Model specific and agnostic methods perform similarily (aside dCAM and XCM).
- xAI methods underperform compared to Full models.
- Causal explanations obtain significantly higher performances.

Équipes Traitement
de l'Information
et Systèmes

Conclusions

First empirical study of explanation quality for relevance attribution maps and causal explanations.

- Post-hoc methods are based on associative mechanisms.
    - Not actionnable (hard to build good surrogate models)
    - Smaller explanation sizes are more consistent and precise.
- In the litterature, trade-off between explanation understandability and explanation performance.
    - Causal notions might go beyond this tradeoff.

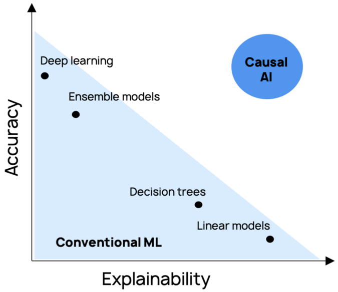


Figure: Explainability tradeoff

Équipes Traitement de l'Information et Systèmes

Future works

- Extend to other explanation methods (prototypes, conceptual, counterfactuals): we need novel metrics to quantify their quality.
- Extend to other prediction tasks and model types (regression, probabilistic).
- Select appropriate explanation methods depending on the dataset

Part II.
Feature selection

Équipes Traitement
de l'Information
et Systèmes

Motivation - Causal feature selection

CONTEXT

MTS data is high-dimensional (lags, number of covariates)

Markov Blanket information is the most relevant to explain a MTS dynamics

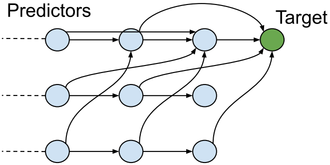
PROBLEM

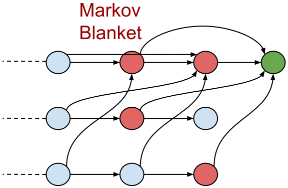
Forecasting model cannot learn on MTS that are too large

Markov Blanket information is generally unavailable in real-world data.

SOLUTION

Causal Feature selection: find the Markov Blanket and exclude other predictors





Context - Traditional approaches

Approaches:

- Filter: Fisher score, Correlation-based, RReliefF
- Information-theory: mRMR selection and/or clustering
- Wrapper: Recursive Feature Elimination
- Embedded: Random Forest gain, LASSO

Downsides:

- Rely on MTS vectorization: transform the considered MTS window into a single dimensional vector. This neglects the temporal structure of the data.
- Only filter approaches scale

Context - Modern approaches

Deep Learning
Temporal Fusion Transformer, Dual-stage Attention-based RNN, Neural Feature selector...
Downsides:

- Curse of dimensionality (problem of embedded methods)
- In most cases, feature selection weights are applied to extracted features.

Causal discovery
PCMCI, SVAR-FCI, Bivariate Granger, SyPI...
Downsides:

- Most techniques are not scalable to high dimensions due to combinatorial explosion
- The scalable Bivariate Granger method is a filter method.

Both approaches are generally not tested for MTS with more than 150 covariates.

Chronomp : high level ideas

The Feature selection problem is *combinatorial*.

We build a feature set in a greedy, heuristic way:

- Forward phase heuristic: start from empty itemset, include variables one by one
- Multivariate: select the variable that explains best the residuals of an intermediate model
- Scalable: use correlation to select variables, to avoid building a model for each variable

Similar approach as Tsagris et al. 2022, Q. Wang and Qin 2013

Algorithm

procedure CHRONOMP(MTS X, target variable i, lags L, stopping threshold α)
    subset of selected features $S \leftarrow \{i\}$
    old model $M' \leftarrow$ None
    new model $M \leftarrow$ MODEL(data: X, target: i, predictors: S, lags: L)
    repeat
        $r \leftarrow$ getResiduals(M)   ▷ vector of size $T - L$
        $selected \leftarrow \arg\max_{c \in X \setminus S}$ ASSOCIATION(r, X^c, L)
        $S \leftarrow S \cup \{selected\}$
        $M' \leftarrow M$
        $M \leftarrow$ MODEL(X, i, S, L)
    until STOPPING-CRITERION(M', M, S, α)
return S or S without last selected item depending on stopping criterion

Algorithm

procedure $\textsc{ChronOMP}$(MTS X, target variable i, lags L, stopping threshold α)
  subset of selected features $S \leftarrow \{i\}$
  old model $M' \leftarrow$ None
  new model $M \leftarrow$ MODEL(data: X, target: i, predictors: S, lags: L)
  repeat
    $r \leftarrow$ getResiduals(M) ▷ vector of size $T - L$
    $selected \leftarrow \arg\max_{c \in X \setminus S}$ ASSOCIATION(r, X^c, L)
    $S \leftarrow S \cup \{selected\}$
    $M' \leftarrow M$
    $M \leftarrow$ MODEL(X, i, S, L)
  until STOPPING-CRITERION(M', M, S, α)
return S or S without last selected item depending on stopping criterion

Équipes Traitement de l'Information et Systèmes

Model component

Auto Regressive Distributed Lags (ARDL) model (Pesaran, Shin, and Smith 2001)

$$y_t = \sum_{j \leq p} a_j . y_{t-j} + \sum_{j \leq p} B_j . X_{t-j} + \sum_{i \leq s} g_i . \mathbb{I}[t = i \mod s] + D.(1, t, t2) + \epsilon_t$$

y_t is the predicted covariate, y_{t-j} its lags, X_{t-j} the matrix of the other covariates at lag j.

- Linear size in function of lags and covariates
- Ordinary Least Squares estimation (fast)
- Robust to heteroscedasticity of noise
- Robust to autocorrelation

Assumptions

- Noise terms are independent
- Stationary, trend-stationary, season-stationary
- Need more observations than there are coefficients in the model (a usual working requirement is 10 times).
- Full rank data
- Finite 4th moment of each covariate

Algorithm

procedure CHRONOMP(MTS X, target variable i, lags L, stopping threshold α)
  subset of selected features $S \leftarrow \{i\}$
  old model $M' \leftarrow$ None
  new model $M \leftarrow$ MODEL(data: X, target: i, predictors: S, lags: L)
  repeat
    $r \leftarrow$ getResiduals(M) ▷ vector of size $T - L$
    $selected \leftarrow \arg\max_{c \in X \setminus S}$ ASSOCIATION(r, X^c, L)
    $S \leftarrow S \cup \{selected\}$
    $M' \leftarrow M$
    $M \leftarrow$ MODEL(X, i, S, L)
  until STOPPING-CRITERION(M', M, S, α)
return S or S without last selected item depending on stopping criterion

Équipes Traitement de l'Information et Systèmes

Association component

Principle: measure the correlation between the residuals and lag 1 to L of a covariate. Then, return the maximal correlation, or the minimal p-value.

Pearson Correlation

- Linear relation
- Interval or ratio-level measurements
- Ideally, data is normally distributed
- No outlier

Spearman Correlation

- Monotonic relation
- Rank-order measurements
- Distribution-free
- No strong outlier

Équipes Traitement de l'Information et Systèmes

Algorithm

procedure CHRONOMP(MTS X, target variable i, lags L, stopping threshold α)
  subset of selected features $S \leftarrow \{i\}$
  old model $M' \leftarrow$ None
  new model $M \leftarrow$ MODEL(data: X, target: i, predictors: S, lags: L)
  repeat
    $r \leftarrow$ getResiduals(M) ▷ vector of size $T - L$
    $selected \leftarrow \arg\max_{c \in X \setminus S}$ ASSOCIATION(r, X^c, L)
    $S \leftarrow S \cup \{selected\}$
    $M' \leftarrow M$
    $M \leftarrow$ MODEL(X, i, S, L)
  until STOPPING-CRITERION(M', M, S, α)
return S or S without last selected item depending on stopping criterion

Équipes Traitement de l'Information et Systèmes

Stopping criterion

procedure STOPPING-CRITERION(current model M, previous model M', selected S, stopping threshold α)
  if $|S| \geq$ max-to-select **then**
    Decision $\leftarrow$ Stop
  if $10 \times |M| \geq |X|$ **then**
    Decision $\leftarrow$ Stop
  if statTest(M, M').pvalue $\leq \alpha$ **then**
    Decision $\leftarrow$ Continue
  else
    Decision $\leftarrow$ Stop
  return Decision

Statistical tests for nested models:

- F-test
- Wald-test
- Likelihood Ratio test

Stopping criterion

F-test	Wald-test	LR-test
Linear model	Linear model (variant for some nonlinear model exist)	Identifiable model by Maximum Likelihood Estimation
Gaussian noise	Asymptotically normal noise in large samples	Asymptotically normal noise in large samples
Independence of noise between timestamps and covariates	Idem	Idem

Table: Stopping criterion assumptions

Équipes Traitement
de l'Information
et Systèmes

Synthetic datasets

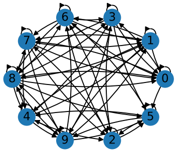


Figure: **Linear dataset VAR10**

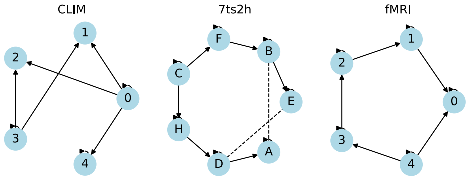


Figure: **Causal datasets CLIM, 7ts2h, fMRI**

- 10 (VAR10) - 10000 (VAR10000) variables, 3500 timesteps
- Generated by a VAR model
- Up to lag 5 dependencies
- Autoregressive component (lags 1 to 5) plus 5 other direct causes (one lag each, between lag 1 and lag 5).

Dataset	Instances	Variables	Timestamps	Avg in-degree	Max lag
CLIM	200	5	250	2.0 to 4.4	2
7ts2h	10	7	4000	1.8	1
fMRI	17	5	200 to 5000	2.0 to 2.6	1

- nonlinearity (7ts2h, fMRI)
- time aggregation (fMRI, CLIM)
- empirical causal graph (CLIM)

Équipes Traitement
de l'Information
et Systèmes

Metrics

We evaluate:

- Predictive performance (**$\mathbf{R^2}$**) of an ARDL model trained on data corresponding to the selected itemset

$$R^2 = 1 - \frac{\text{sum of squared residuals}}{\text{total variance of the predicted variable}}$$

- Overlap of the set of selected covariate S with the set of direct causes C:
  - **Recall**: how much of the direct causes are selected
  - **Precision**: how much of the selected variables are direct causes.
  - Recall is especially important, as we primarily seek to select all relevant variables.

$$\text{Recall} = \frac{|S \cap C|}{C} \qquad \text{Precision} = \frac{|S \cap C|}{S}$$

Results - VAR data - R2

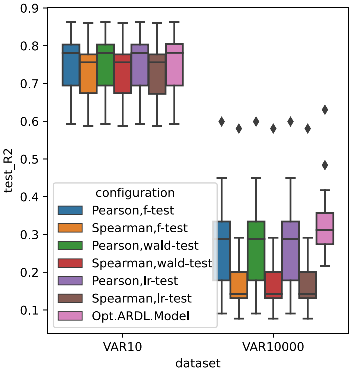


Figure: R2 of each configuration

Observations

- On linear data, Pearson correlation is better suited to the ARDL model
- The kind of stopping criterion has low impact on the final R^2
- The Pearson correlation configurations are close to the optimal ARDL model (pink boxes)

Results - VAR data - Recall, Precision

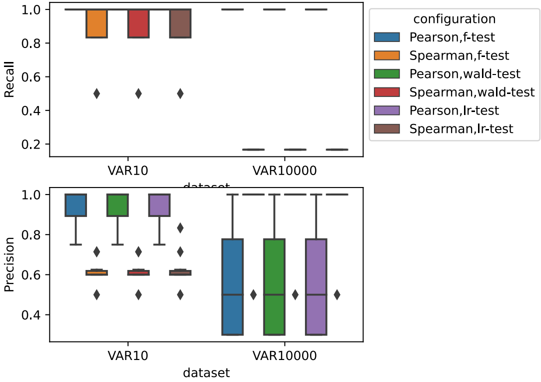


Figure: Recall of each FS configuration

Observations

- Pearson configurations select all direct causes
- Spearman misses most causes in large data
- Spearman high precision in VAR10000 indicate early stopping, while Pearson low precision indicate late stopping

Équipes Traitement
de l'Information
et Systèmes

Results - VAR data - Graph building

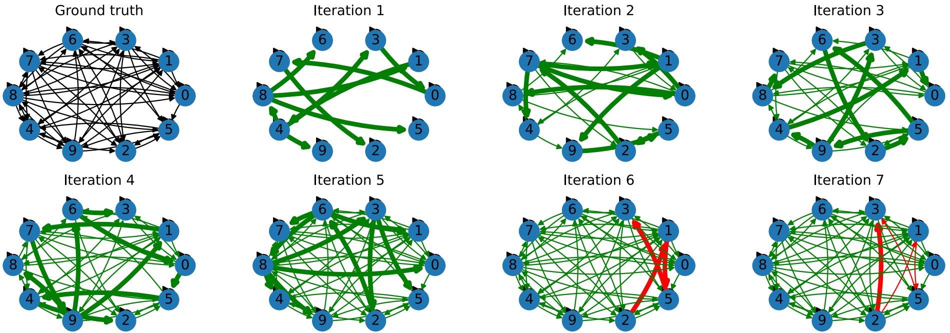


Figure: Evolution of graph built for all 10 variables, at each iteration of the algorithm. Green: direct causes, Red: other covariates, **Bold**: added at current step

Équipes Traitement
de l'Information
et Systèmes

Results - VAR data - Time

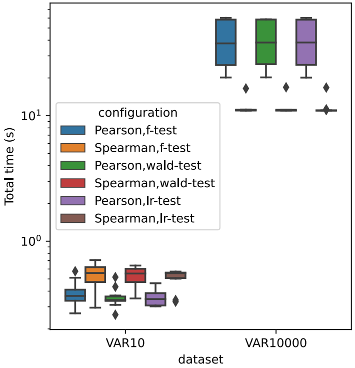


Figure: Computing time (s) of each FS configuration

The difference in time spent can be explained by the differing number of iterations. For instance, on VAR10000, Spearman configurations select less variables than Pearson configurations.

Results - Causal Discovery data - R2

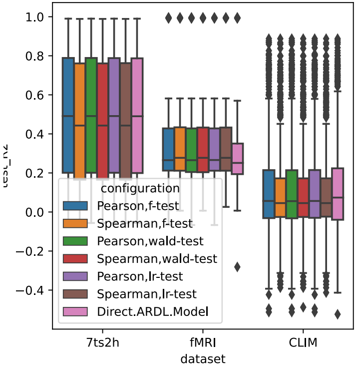


Figure: R2 of each configuration

Observations

- The kind of stopping criterion has low impact on the final R^2
- Pearson is better on 7ts2h, CLIM, while Spearman on fMRI
- On nonlinear data, direct causes might not necessarily be the best features for a linear model
- The Pearson correlation configurations are closer to the ARDL model built on direct causes (pink boxes)

Results - Causal discovery data - Recall, Precision

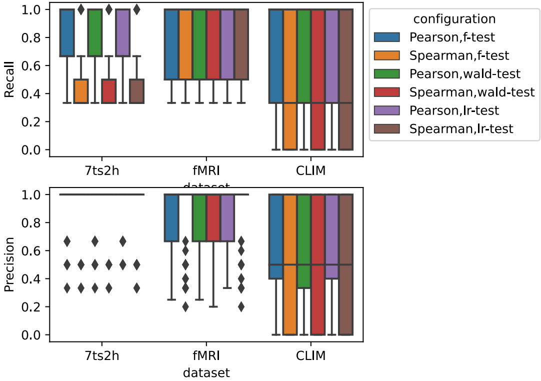


Figure: Recall of each FS configuration

Observations

- Pearson configurations median recall is 1
- On fMRI, Spearman,f-test and Spearman,wald-test has a small edge in precision

Baselines

Atemporal: **Recursive Feature Elimination** (Guyon et al. 2002)

- Backward selection method
- Vectorize time series
- Returns pairs (variable, lag)
- User-specified number of features (here set as the number of causes)
- Not scalable (wrapper approach)
- Wrapper FS: multivariate interactions

Choice made: transform the returned itemset from pairs (variable, lag) to singleton (variable,). Compute precision, recall on this itemset, compute R^2 on an ARDL model for this itemset.

Temporal: **Bivariate Granger** (Sun et al. 2015)

Algorithm:

- Fit VAR models on each pair (target, covariate)
- Test causality from each covariate to the target
- Select covariates that cause the target but not the opposite

Properties:

- Linear interactions only
- Scalable (linear in number of covariates)
- Filter FS: bivariate interactions only

Baselines - Results

	7ts2h	CLIM	VAR10	fMRI	VAR10000
bg	0.489999	0.102695	0.495436	0.351780	NaN
rfe	0.491199	**0.108432**	0.660708	0.340141	NaN
chronomp	**0.493559**	0.105842	**0.747287**	**0.357854**	**0.273421**

Table: Average R2 of ARDL model on selected features

	7ts2h	CLIM	VAR10	fMRI	VAR10000
ground truth	2.4	1.5	6	2.0	6
bg	2.142857	1.644031	1.133333	1.435294	484.70
rfe	1.828571	2.044907	4.000000	1.458824	NaN
chronomp	1.900000	1.685652	6.400000	1.623529	13.35

Table: Average size of selected set

- chronomp dominates on VAR10
- Similar R2 in other small datasets
- BG and RFE aren't practical for VAR10000

Équipes Traitement
de l'Information
et Systèmes

Baselines - Recall, Precision

	7ts2h	CLIM	VAR10	fMRI	VAR10000
bg	0.730952	0.651406	0.188889	0.600000	0.575
rfe	0.745238	**0.881763**	0.666667	0.625490	NaN
chronomp	**0.747619**	0.670938	**1.000000**	**0.643137**	**1.000000**

Table: Average recall over the direct causes

	7ts2h	CLIM	VAR10	fMRI	VAR10000
bg	0.857381	0.613874	**1.000000**	0.871569	0.007119
rfe	**0.940476**	**0.735542**	**1.000000**	**0.873529**	NaN
chronomp	0.926190	0.619095	0.946429	0.849804	**0.544552**

Table: Average precision over the direct causes

- chronomp dominates on VAR
- chronomp recall higher otherwise

- RFE dominates on CLIM (translating to 3% increase in R^2)

Équipes Traitement de l'Information et Systèmes

Baselines - Results

	7ts2h	CLIM	VAR10	fMRI	VAR10000
bg	0.092009	**0.017455**	**0.139232**	**0.024697**	134.649533
rfe	28.721915	0.026837	87.259455	2.465924	NaN
chronomp	**0.084844**	0.026936	0.385377	0.039340	**41.073843**

Table: Average time spent (s) in FS phase

Observations

- RFE method sklearn implementation scales badly with number of observations
- chronomp takes longer than BG, but stays within similar magnitudes.

Équipes Traitement de l'Information et Systèmes

Conclusion and future works

Main conclusion:

- Promising results (high recall)
- Robust to high numbers of covariates

Improvements to be made:

- Include backward phase to increase Precision
- Explore branch-and-bound to create equivalent feature sets
- Use constrained linear models for data parsimony and time efficiency.

Experimental setup to put in place:

- Standardize R^2 evaluation with Auto-ML selection of best final model for each FS method.
- Improve dataset diversity (real datasets, large datasets, nonlinear datasets, lag diversity)

Équipes Traitement
de l'Information
et Systèmes

Laboratoire ETIS
6 Avenue du Ponceau
95000 Cergy

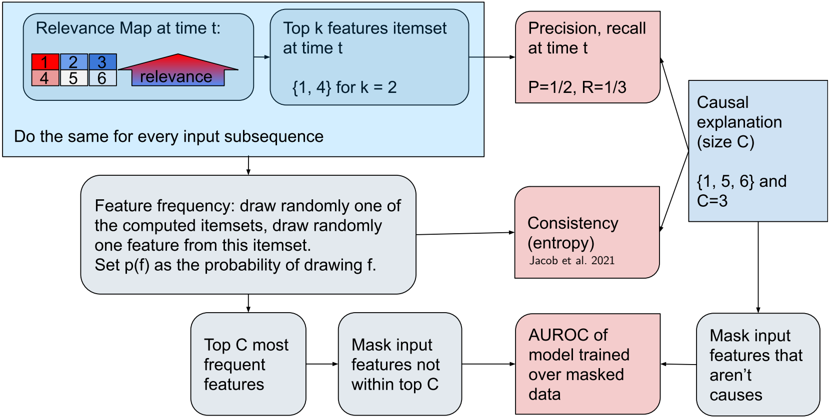
T.07 61 76 91 47

etis-lab.fr

Thank you for your attention!

Équipes Traitement
de l'Information
et Systèmes

More about metrics



Bibliography I

- Assaad, Charles K., Emilie Devijver, and Éric Gaussier (2022). "Survey and Evaluation of Causal Discovery Methods for Time Series". In: *J. Artif. Intell. Res.* 73, pp. 767–819.
- Bodria, Francesco et al. (2021). "Benchmarking and Survey of Explanation Methods for Black Box Models". In: *CoRR*.
- causalens.com. URL: https://www.causalens.com/blog/xai-doesnt-explain/.
- Guyon, Isabelle et al. (2002). "Gene Selection for Cancer Classification Using Support Vector Machines". In: *Mach. Learn.* 46.1–3, pp. 389–422.
- Huang, Yuxiao and Samantha Kleinberg (2015). "Fast and Accurate Causal Inference from Time Series Data.". In: *FLAIRS Conference*, pp. 49–54.
- Ismail, Aya Abdelsalam et al. (2020). "Benchmarking Deep Learning Interpretability in Time Series Predictions". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., pp. 6441–6452.

Équipes Traitement
de l'Information
et Systèmes

Bibliography II

Jacob, Vincent et al. (2021). “Exathlon: A Benchmark for Explainable Anomaly Detection over Time Series”. In: *Proc. VLDB Endow.* 14.11, pp. 2613–2626. ISSN: 2150-8097. DOI: 10.14778/3476249.3476307.

Nauta, Meike et al. (2023). “From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI”. In: *ACM Comput. Surv.* ISSN: 0360-0300. DOI: 10.1145/3583558. URL: https://doi.org/10.1145/3583558.

Pesaran, M. Hashem, Yongcheol Shin, and Richard J. Smith (2001). “Bounds testing approaches to the analysis of level relationships”. In: *Journal of Applied Econometrics* 16.3.

Q. Wang, X. Li and Q. Qin (2013). “”Feature Selection for Time Series Modeling””. In: *Journal of Intelligent Learning Systems and Applications* 5.3.

Équipes Traitement
de l'Information
et Systèmes

Bibliography III

Runge, Jakob et al. (2020). "The Causality for Climate Competition". In: *Proceedings of the NeurIPS 2019 Competition and Demonstration Track*. Ed. by Hugo Jair Escalante and Raia Hadsell. Vol. 123. Proceedings of Machine Learning Research. PMLR, pp. 110–120.

Sun, Youqiang et al. (2015). "Using Causal Discovery for Feature Selection in Multivariate Numerical Time Series". In: *Mach. Learn.* 101.1–3, pp. 377–395.

Tsagris, M. et al. (2022). "The γ -OMP Algorithm for Feature Selection With Application to Gene Expression Data". In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 19.02, pp. 1214–1224.