Imperial College London

3rd November 2021

Causal Injection into Neural Networks

ACM ICAIF'21: Workshop on XAI in Finance

Fabrizio Russo & Francesca Toni

Imperial College London

Outline

Introduction & Background

Motivation

- In finance many hard problems are tackled with models (e.g. fraud, pricing, credit scoring, trading, planning etc.)

Imperial College London

Motivation

- In finance many hard problems are tackled with models (e.g. fraud, pricing, credit scoring, trading, planning etc.)
- Practitioners often have a lot of domain (causal) knowledge

Imperial College London

Motivation

- In finance many hard problems are tackled with models (e.g. fraud, pricing, credit scoring, trading, planning etc.)
- Practitioners often have a lot of domain (causal) knowledge
- Regulation is quite strict in requiring model stakeholders to understand and “own” their models

Imperial College London

Motivation

- In finance many hard problems are tackled with models (e.g. fraud, pricing, credit scoring, trading, planning etc.)
- Practitioners often have a lot of domain (causal) knowledge
- Regulation is quite strict in requiring model stakeholders to understand and “own” their models
- Machine Learning models (e.g. Neural Networks) do not easily allow knowledge integration nor interpretation

Motivation

- In finance many hard problems are tackled with models (e.g. fraud, pricing, credit scoring, trading, planning etc.)
- Practitioners often have a lot of domain (causal) knowledge
- Regulation is quite strict in requiring model stakeholders to understand and "own" their models
- Machine Learning models (e.g. Neural Networks) do not easily allow knowledge integration nor interpretation

Causal Injection into Neural Networks

Introducing causality into neural networks not only makes them more **robust and reliable**, but it is also a step towards their **interpretability**.

Imperial College London

Formal Set-up

- Let $X_1, \ldots, X_d$ be the set of *input features* and Y be the *target feature* within a regression or classification setting
  - $f_Y : \mathcal{X} \rightarrow \mathcal{Y}$

Formal Set-up

- Let $X_1, \dots, X_d$ be the set of *input features* and Y be the *target feature* within a regression or classification setting
  - $f_Y : \mathcal{X} \to \mathcal{Y}$
- Causal Structure is a DAG $\mathcal{G} = \langle V, E \rangle$ (Pearl 2009)
  - $V = \{Y, X_1, \dots, X_d\}$ the set of vertices
  - $E \subseteq V \times V$ the set of edges

Imperial College London

Formal Set-up

- Let $X_1, \ldots, X_d$ be the set of *input features* and Y be the *target feature* within a regression or classification setting
  - $f_Y : \mathcal{X} \to \mathcal{Y}$
- Causal Structure is a DAG $\mathcal{G} = \langle V, E \rangle$ (Pearl 2009)
  - $V = \{Y, X_1, \ldots, X_d\}$ the set of vertices
  - $E \subseteq V \times V$ the set of edges
- $v_i = f_i(pa_i, u_i)$
  - v_i is a value for $V_i \in V$ with parents Pa_i having values pa_i
  - f_i any function
  - u_i representing the errors due to omitted factors

Background

- Causal Structure Learning (CASTLE by Kyono, Zhang and Schaar 2020) use a causal discovery method to regularise neural networks
  - A *joint* neural network learns the causal DAG underpinning the data as an adjacency matrix while predicting / reconstructing every feature

Background

- Causal Structure Learning (CASTLE by Kyono, Zhang and Schaar 2020) use a causal discovery method to regularise neural networks
  - A *joint* neural network learns the causal DAG underpinning the data as an adjacency matrix while predicting / reconstructing every feature
- Issue is: CASTLE prefers using parents to children and siblings, but it is not **guaranteed** to do so.

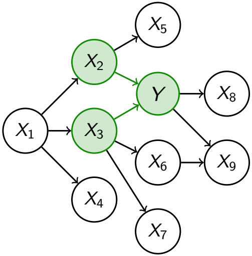
Background

- Causal Structure Learning (CASTLE by Kyono, Zhang and Schaar 2020) use a causal discovery method to regularise neural networks
    - A *joint* neural network learns the causal DAG underpinning the data as an adjacency matrix while predicting / reconstructing every feature
- Issue is: CASTLE prefers using parents to children and siblings, but it is not **guaranteed** to do so.

Can we **make sure** a neural network **complies** with a given DAG?

Imperial College London

Synthetic Data Example



(a)

	Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
Y	0.0	0.005	0.017	0.008	0.002	0.042	0.02	0.005	0.059	0.05
X_1	0.006	0.0	0.063	0.054	0.068	0.009	0.006	0.013	0.006	0.008
X_2	0.088	0.036	0.0	0.022	0.019	0.124	0.008	0.011	0.006	0.008
X_3	0.087	0.034	0.021	0.0	0.024	0.005	0.107	0.104	0.006	0.009
X_4	0.009	0.032	0.02	0.023	0.0	0.01	0.013	0.01	0.005	0.005
X_5	0.026	0.006	0.017	0.004	0.004	0.0	0.012	0.002	0.005	0.018
X_6	0.025	0.006	0.008	0.011	0.005	0.017	0.0	0.014	0.002	0.114
X_7	0.029	0.003	0.007	0.011	0.002	0.024	0.029	0.0	0.011	0.01
X_8	0.036	0.002	0.004	0.003	0.004	0.006	0.009	0.006	0.0	0.006
X_9	0.024	0.003	0.003	0.004	0.003	0.005	0.079	0.01	0.004	0.0

(b) $w_{ik} = \sqrt{\sum_{j=1}^{h} \left(\Theta_1^{i,j,k}\right)^2}$

Figure 1: (a) Example DAG from Kyono, Zhang and Schaar 2020.
(b) Adjacency Matrix produced by CASTLE when fitted to the synthetic data produced following the DAG to the left.

Algorithm 1 - Inject Causal Knowledge

Imperial College London

The Intuition

- **Objective:** have the network use only the relationships contained in the DAG i.e. predict each of the features using only its parents.

	Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
Y	0.0	0.005	0.017	0.008	0.002	0.042	0.02	0.005	0.059	0.05
X_1	0.006	0.0	0.063	0.054	0.068	0.009	0.006	0.013	0.006	0.008
X_2	0.088	0.036	0.0	0.022	0.019	0.124	0.008	0.011	0.006	0.008
X_3	0.087	0.034	0.021	0.0	0.024	0.005	0.107	0.104	0.006	0.009
X_4	0.009	0.032	0.02	0.023	0.0	0.01	0.013	0.01	0.005	0.005
X_5	0.026	0.006	0.017	0.004	0.004	0.0	0.012	0.002	0.005	0.018
X_6	0.025	0.006	0.008	0.011	0.005	0.017	0.0	0.014	0.002	0.114
X_7	0.029	0.003	0.007	0.011	0.002	0.024	0.029	0.0	0.011	0.01
X_8	0.036	0.002	0.004	0.003	0.004	0.006	0.009	0.006	0.0	0.006
X_9	0.024	0.003	0.003	0.004	0.003	0.005	0.079	0.01	0.004	0.0

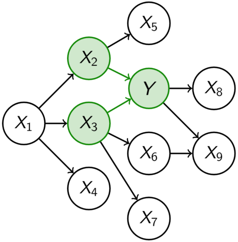
	Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
Y	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.059	0.05
X_1	0.0	0.0	0.063	0.054	0.068	0.0	0.0	0.0	0.0	0.0
X_2	0.088	0.0	0.0	0.0	0.0	0.124	0.0	0.0	0.0	0.0
X_3	0.087	0.0	0.0	0.0	0.0	0.0	0.107	0.104	0.0	0.0
X_4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
X_5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
X_6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.114
X_7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
X_8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
X_9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Figure 2: Enforce acyclicity and threshold on CASTLE adjacency matrix:

$$E_\tau(\mathbf{W}) = \{(i,k) | w_{ik} > w_{ki} \wedge w_{ik} > \tau\}$$

The Intuition

- **Objective:** have the network use only the relationships contained in the DAG i.e. predict each of the features using only its parents.

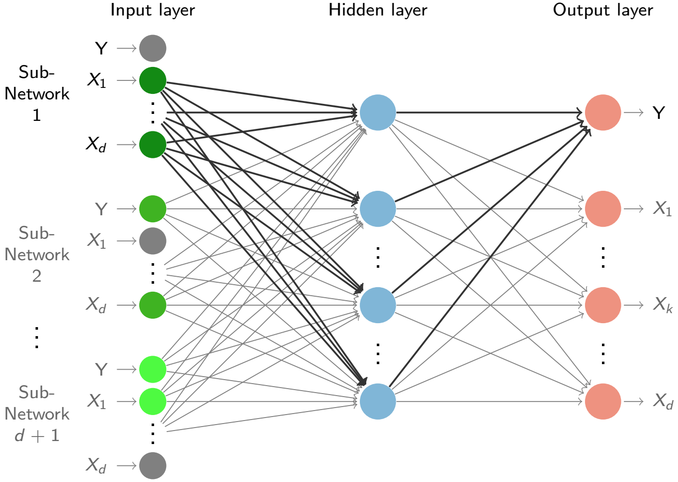


	Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
Y	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.059	0.05
X_1	0.0	0.0	0.063	0.054	0.068	0.0	0.0	0.0	0.0	0.0
X_2	0.088	0.0	0.0	0.0	0.0	0.124	0.0	0.0	0.0	0.0
X_3	0.087	0.0	0.0	0.0	0.0	0.0	0.107	0.104	0.0	0.0
X_4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
X_5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
X_6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.114
X_7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
X_8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
X_9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Figure 3: Enforce acyclicity and threshold on CASTLE adjacency matrix:

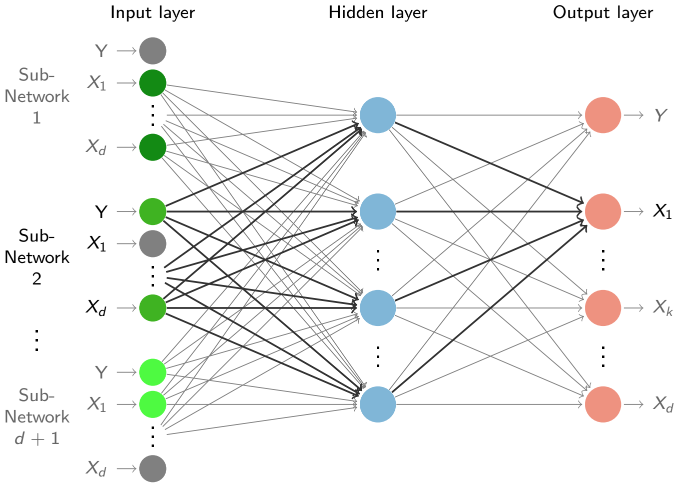
$$E_\tau(\mathbf{W}) = \{(i,k) | w_{ik} > w_{ki} \wedge w_{ik} > \tau\}$$

Joint Network Structure



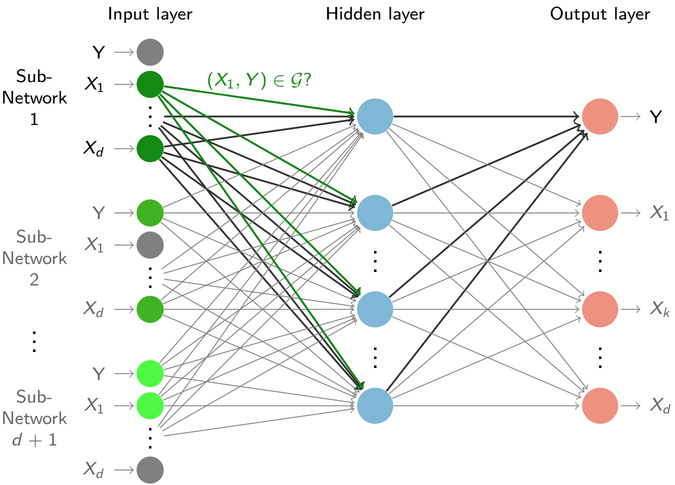
Imperial College London

Joint Network Structure



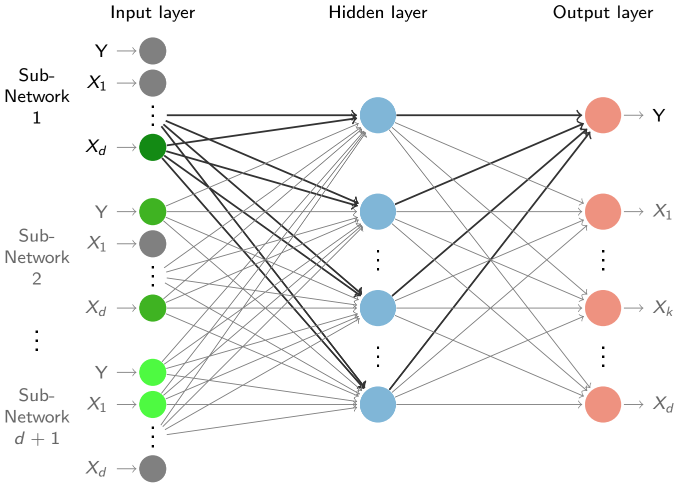
Imperial College London

Algorithm 1 - Inject Causal Knowledge



Imperial College London

Algorithm 1 - Inject Causal Knowledge



Imperial College London

Algorithm 1 - Limitations & Opportunities

- It requires a complete DAG (covering all variables considered in the problem and the data)

Algorithm 1 - Limitations & Opportunities

- It requires a complete DAG (covering all variables considered in the problem and the data)
- Full causal DAG is rare and often impractical to build

Algorithm 1 - Limitations & Opportunities

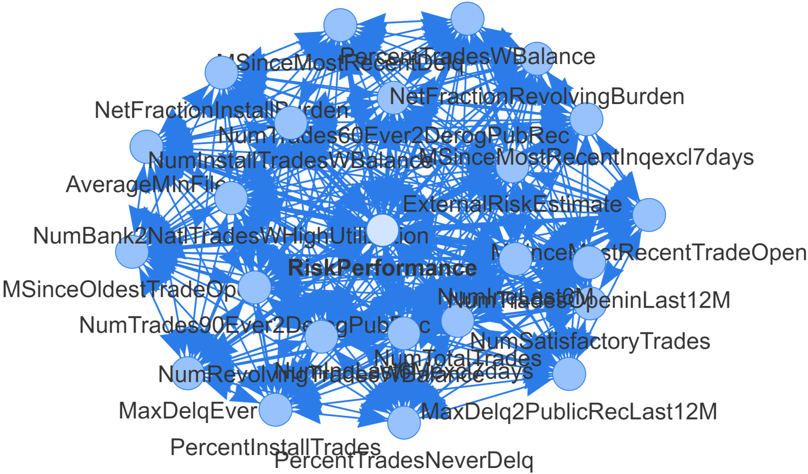
- It requires a complete DAG (covering all variables considered in the problem and the data)
- Full causal DAG is rare and often impractical to build
- We propose a second algorithm that involves Subject Matter Experts (SMEs) providing their input

Algorithm 2 - Refine & Inject DAG: A Credit Risk Case Study

FICO/HELOC dataset

- Public *credit risk* dataset from a challenge on explainable ML (FICO 2017).
- 10k observation and 24 features.
- Target Y is the *RiskPerformance* metric: whether a debtor has always paid their dues for the two years after being granted a loan.
- Features include the usual ones e.g. credit score, payment history, search history etc.

Starting Point - CASTLE



Explore Different DAGs

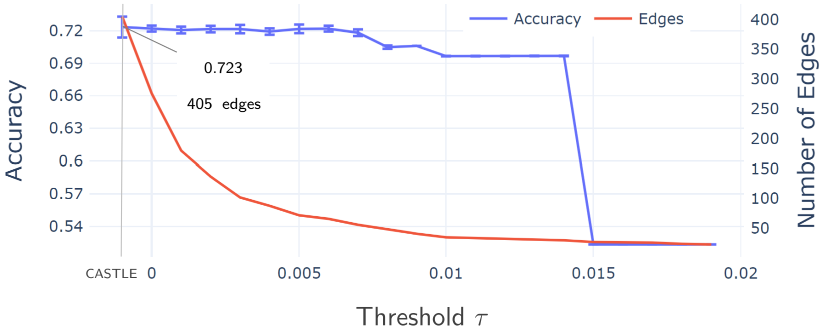


Figure 4: Change in accuracy and number of edges in the DAG when changing the threshold τ. $E_\tau(\mathbf{W}) = \{(i,k) | w_{ik} > w_{ki} \wedge w_{ik} > \tau\}$

Imperial College London

Explore Different DAGs

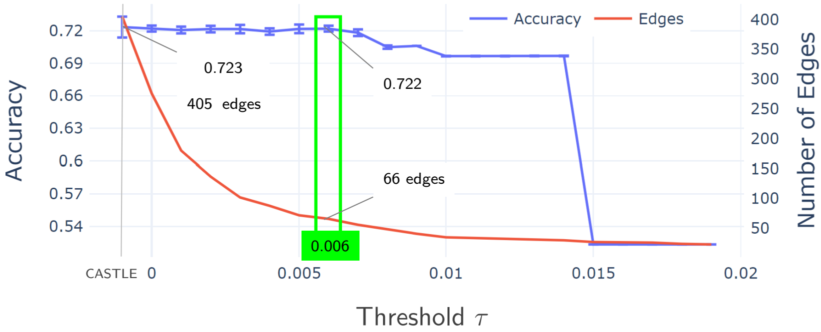
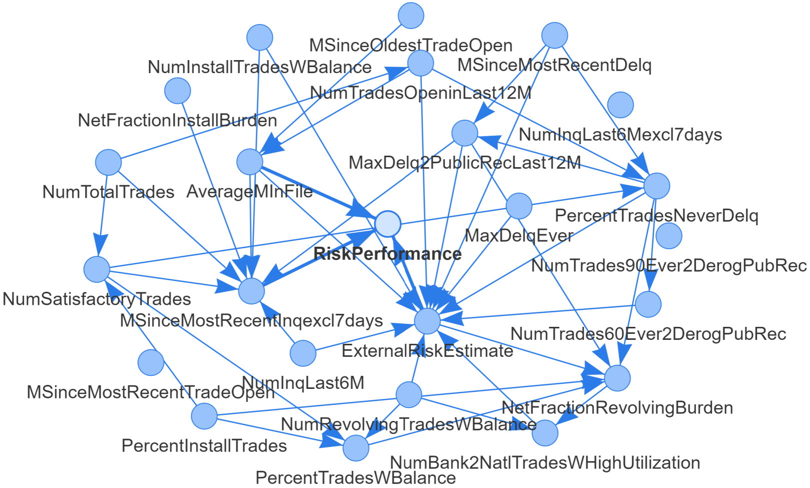


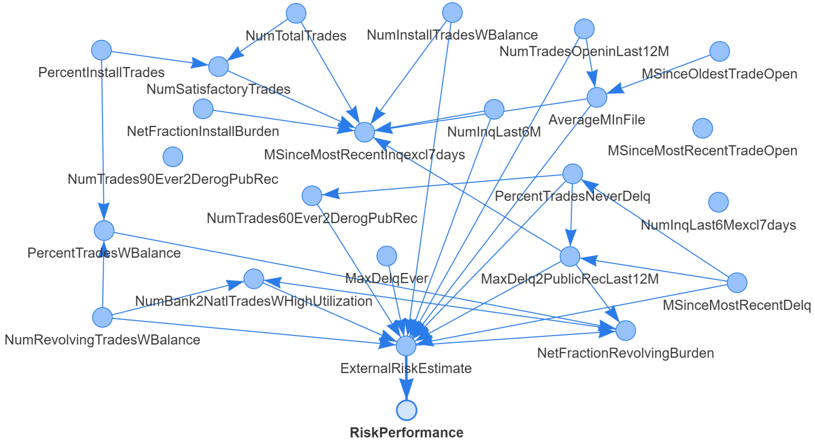
Figure 4: Change in accuracy and number of edges in the DAG when changing the threshold τ. $E_\tau(\mathbf{W}) = \{(i,k)|w_{ik} > w_{ki} \wedge w_{ik} > \tau\}$

Imperial College London

Initial Refinement - $\tau = 0.006$



Build DAG Bottom up - $\tau = 0.012$



Conclusion

Imperial College London

Conclusion

We showed:

- how to introduce causal representation guarantees by making a neural network adhere to an input causal DAG
- that causal injection can drastically reduce the amount of weights in a network while
  - maintaining comparable performance
  - improving robustness and interpretability

Thank You

Questions?

References I

FICO (2017). *FICO xML Challenge found at community.fico.com/s/xml*. URL: https://community.fico.com/s/explainable-machine-learning-challenge.

Kyono, Trent, Yao Zhang and Mihaela van der Schaar (2020). 'CASTLE: Regularization via Auxiliary Causal Graph Discovery'. In: *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*. URL: https://proceedings.neurips.cc/paper/2020/hash/1068bceb19323fe72b2b344ccf85c254-Abstract.html.

Pearl, Judea (2009). *Causality*. Cambridge university press.