SAM

Structural Agnostic Model, causal discovery and penalized adversarial learning

D.Kalainathan, O.Goudet, D.Lopez-Paz, P.Caillou, I.Guyon, M.Sebag

TAU, CNRS, INRIA,
Université Paris Sud,
Université Paris Saclay, France
Facebook AI Research

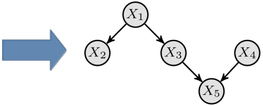
Outline of the talk

Introduction

Causal Discovery

- Goal: build a graph which models how the data could have been generated

X1	X2	X3	X4	X5
-0.5862258940391734	0.37081241324802505	-0.6183613167803281	-1.482229977659025	0.9680486293593741
0.38941645291310745	-0.9331472003238432	-0.9100680407589063	2.25960201340443	-0.42257395495166383
-0.20220983054366237	-0.44517749475963164	0.34037124141265496	0.11466803735056992	-0.06948041619106957
-0.7720994805068446	-0.8091374447229811	0.3449576991826806	0.34858920665874743	0.6522438917449829
-0.010768921199488801	-0.37216622644182973	-0.05464190150369652	0.13037211033392554	-0.4913403880980598
0.5664215775261501	-0.1049209674314886	-0.46873511475192275	0.6255101327326772	-1.1161363764783467
0.07609329697595602	0.4349222132318391	0.03616237121000654	-0.5851187482036869	-0.15094649255263504
-0.8105014020060939	-0.85009038661454	0.08297081861478284	0.4219202849261392	1.2083824842312032
0.26960803769603325	-0.5457664267582014	-0.2399565515113291	0.9930418368462299	-0.9952398322884214
-0.2534659693146175	-0.1096282972825271	0.4194935296222515	-0.6831858701108589	-0.3332932450733603
0.3726333036636913	-0.5474835352715373	-0.42134231750637374	1.0369903645998773	-0.9932358201077405
-0.7440696522505421	0.30540051210870156	0.7651736196873122	-1.3878221578400844	1.429932163656556
0.41574261047898964	-0.346746665073004	-0.22928220688004908	0.7806010297284206	-0.7122954369831308
-0.5183600813593083	0.2992187501306689	-0.07916853473166507	-1.4399861909706322	1.2141727308614034
-0.09549337432227678	-0.616677040864746	-0.7888041661301364	0.6530364786439687	-0.1575425056756211
0.08425731144796944	0.0518242083304139	-0.29936719343655205	-0.46750021903217553	-0.3604728034952744
0.6187635111404841	0.6602807590674153	0.31871729174162655	0.26294545150556964	0.18583704209040075
0.13356931819000462	0.9084817849462865	-0.8061598179894509	-0.17123037653356019	0.51763823098546
-0.6212547839216225	0.47365766778948015	-0.9670935473077173	-1.5882484961180763	2.201232674683203
-0.4932640158865276	0.5673686445337787	0.6767890869554347	-1.2720236083602237	2.752251581397096
-0.36767523892558507	-0.7313388799818274	0.6756014545748446	0.5110727422888668	0.5715208150253143
-0.17317926886018453	0.8497898555599273	-0.4335031512208065	-0.631441611637212	-0.03952023419554657
-0.47476120454940807	-0.1273390262809646	-0.6741159352826112	-1.086198238062107	-0.4291672030734802
-0.6126586955166871	-0.23508888658747895	0.4067451595829741	-0.7633616433451832	0.006554869926958593
0.6706332542475919	-0.2572601710935667	-0.820552294335464	1.1080816400161335	-1.167241493061066
-0.9052815353648547	0.2611922811702674	-0.8720553799118256	-1.7400556373631628	2.2712956130801802
0.5744048928587386	-0.7330300220050077	0.3235838886516609	1.865171012717555	-0.906901640387717
-0.5279998019429426	-0.702142969638422	-0.6999048993995027	0.31091725940062764	0.22665202467072665



- Gold Standard: perform randomized experiments
- Our setting: having only observational data: infer causal relationships in the dataset.

Correlation does not imply causation

- Task: predict a target variable Y given (X_1, X_2)
- Generative process underlying (X_1, X_2, Y):

$$X_1, E_{X_1}, E_{X_2} \sim \text{Uniform}(0,1), X_1 \perp\!\!\!\perp E_{X_1},\ \ Y \perp\!\!\!\perp E_{X_2}$$
$$Y \leftarrow 0.5X_1 + E_{X_1},$$
$$X_2 \leftarrow Y + E_{X_2},$$

$$X_2 \leftarrow Y \leftarrow X_1$$

- Least-squares solution: $\hat{Y} = 0.25X_1 + 0.5X_2$
- X_2 is a better predictor for Y than X_1
- However X_2 does not cause Y

State of the art

Causality - Key idea 1 : Identifiying v-structures

Exploit conditional independence to identify causal relations [Spirtes et al., 2000], [Tsamardinos et al., 2006]:

Skeleton: $A - B - C$

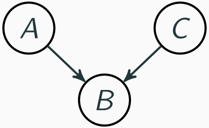
3 Markov Equivalent Classes: $A \perp\!\!\!\perp C | B$

$A \rightarrow B \rightarrow C$ $\qquad$ $A \leftarrow B \leftarrow C$ $\qquad$ $A \leftarrow B \rightarrow C$

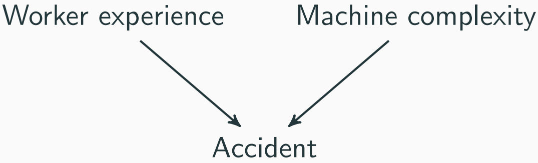
Example

$$\begin{tikzcd}
 & \text{Country's wealth} \arrow[dl] \arrow[dr] & \\
\text{Chocolate consumption} & & \text{Number of Nobel prizes}
\end{tikzcd}$$

V-Structure: $A \not\perp\!\!\!\perp C | B$

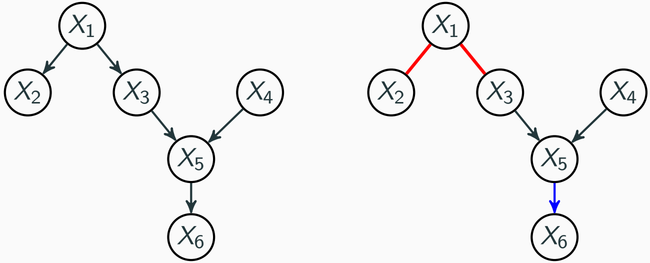


Example



Constraint-based methods

Constraint-based methods, through V-Structures and constraint propagation, output a **CPDAG** (Completed Partially Directed Acyclic Graph). PC algorithm [Spirtes et al., 2000].



(a) The exact DAG of $\mathcal{G}$. (b) The CPDAG of $\mathcal{G}$.

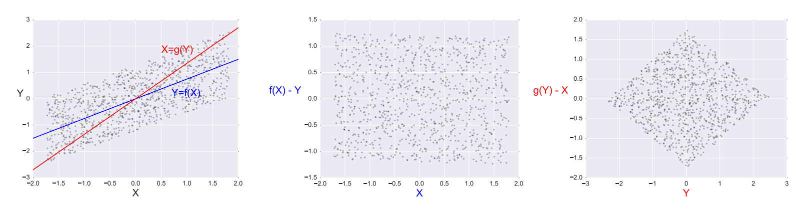
Limitations

- Explosive number of conditional independence tests to perform
- Cannot identify all cause effect relations

$$
\begin{aligned}
X_1, E_{X_1}, E_{X_2} \sim \text{Uniform}(0,1), X_1 \perp\!\!\!\perp E_{X_1},\ \ Y \perp\!\!\!\perp E_{X_2} \\
Y \leftarrow 0.5X_1 + E_{X_1}, \\
X_2 \leftarrow Y + E_{X_2},
\end{aligned}
$$

- Here $X_1 \perp\!\!\!\perp X_2 | Y$. No V-structure

Key idea 2: exploit asymmetry between cause and effect



- Causal additive noise model (ANM) [Hoyer et al., 2009]: $Y = f(X) + E$, with $X \perp\!\!\!\perp E$
- Perform a regression and check independence of the residual and the cause

- With the previous example and a linear causal additive noise model:

$$
\begin{aligned}
X_1, E_{X_1}, E_{X_2} \sim \mathrm{Uniform}(0,1), X_1 \perp\!\!\!\perp E_{X_1},\ Y \perp\!\!\!\perp E_{X_2} \\
Y \leftarrow 0.5X_1 + E_{X_1}, \\
X_2 \leftarrow Y + E_{X_2},
\end{aligned}
$$

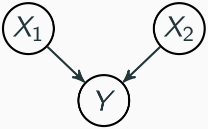
- Regression Y on X_1: $X_1 \rightarrow Y$
- Regression X_2 on Y: $Y \rightarrow X_2$

$$X_2 \leftarrow Y \leftarrow X_1$$

Limitation

- Generality problem: sometimes the causal additive noise model does not hold in any direction
  e.g.: does not work for $Y = X \times E$
- Do not take into account independence relations. Consider:

$$X_1, X_2, E_{X_1} \sim \text{Gaussian}(0,1), X_1 \perp\!\!\!\perp E_{X_1},\ X_2 \perp\!\!\!\perp E_{X_1}$$
$$Y \leftarrow 0.5X_1 + X_2 + E_{X_1}$$



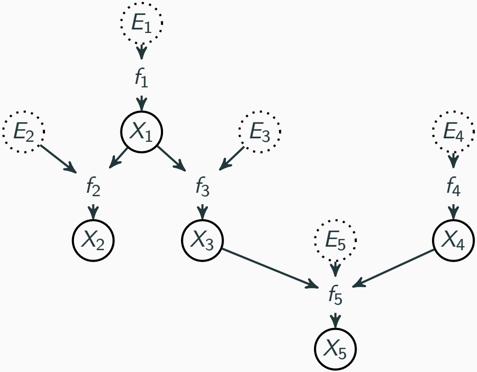
- (X_1, Y) and (X_2, Y) are perfect symmetric pairwise distribution (after rescaling)
- However $X_1 \not\perp\!\!\!\perp X_2 | Y$. A V-structure may be identified

Learning functional Causal Models with generative neural networks

Functional Causal Models (FCMs)

$$X_i = f_i(X_{\mathrm{Pa}(i;\mathcal{G})}, E_i), \forall i \in [1, d]$$

$X_{\mathrm{Pa}(i;\mathcal{G})}$ the set of parents of X_i in $\mathcal{G}$, E_i a random independent noise variable, f_i a deterministic function



$$\begin{cases} X_1 = f_1(E_1) \\ X_2 = f_2(X_1, E_2) \\ X_3 = f_3(X_1, E_3) \\ X_4 = f_4(E_4) \\ X_5 = f_5(X_3, X_4, E_5) \end{cases}$$

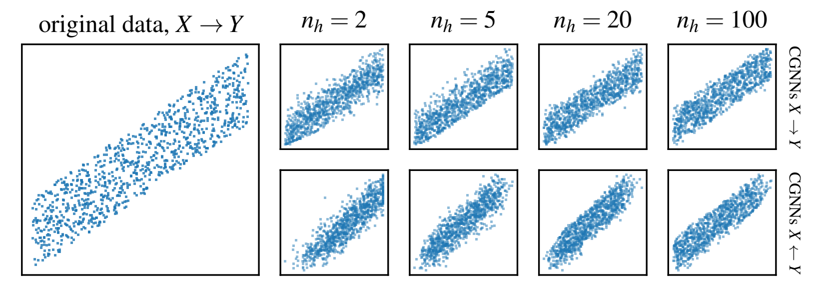
Modelling FCM with generative neural network

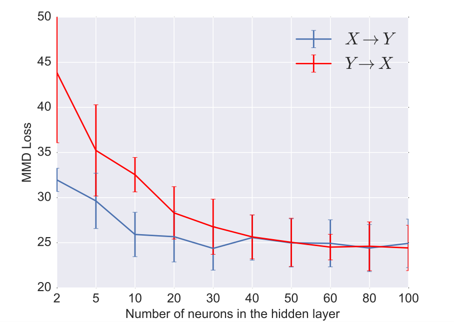
- Idea: approximate the continuous mechanism $f_1, \ldots, f_d$ with a set of one hidden layer neural networks $\hat{f} = (\hat{f}_1, \ldots, \hat{f}_d)$
- Estimate FCMs C as $\hat{C} = (\hat{\mathcal{G}}, \hat{f}, \hat{Q})$:

$$\hat{X}_i \leftarrow \hat{f}_i(\hat{X}_{\mathrm{Pa}(i;\hat{\mathcal{G}})}, \hat{E}_i), \hat{E}_i \sim \hat{Q}, \tag{1}$$

- We can draw a sample $\hat{x} = (\hat{x}_1, \ldots, \hat{x}_d)$ from the distribution $\hat{P} := \hat{P}(X)$:
  1. Draw $\hat{e}_i \sim \hat{Q}$ for all $i = 1, \ldots, d$.
  2. Construct $\hat{x}_i = \hat{f}_i(\hat{x}_{\mathrm{Pa}(i;\mathcal{G})}, \hat{e}_i)$ in the topological order of $\hat{\mathcal{G}}$

Complexity/reproduction trade-off





How to deal with the complexity/reproduction trade-off ?

Two ideas:

1. *Wrapper* approach
   - CGNN - `https://arxiv.org/abs/1711.08936` (XCVML 2018)
   - Limit the number of available hidden units n_h
   - Explore the space of possible graph and find the DAG minimizing a MMD score
2. *Embedded* approach
   - SAM: Structural Agnostic Model
   - Use regularization to enforce automatically the sparsity of the graph and to choose the orientation of the edges
   - Use GAN as a score to reproduce the distributions

SAM: Redefining causal graph recovery as a single optimization problem

Motivation

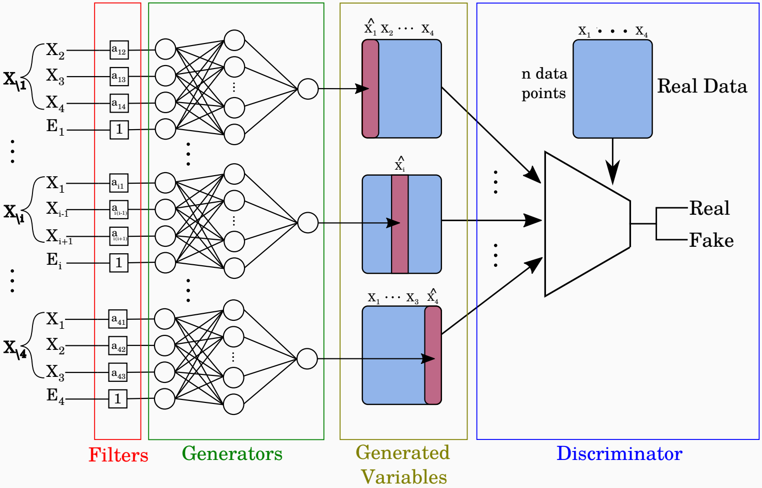
$\rightarrow$ **Minimizing both structural and functional complexity of the graph**

Compared to score-based methods:

- For traditional score based methods, the graph structure search in the graph space (2^d given a skeleton)
- Retraining fully a regression model for each candidate graph

$$X_i = f_i(X_{\setminus i}, E_i), \tag{2}$$

Model Diagram



$$\hat{X}_i = m_i^\top \tanh\left(\bar{W}_i^\top (a_i \odot X) + n_i E_i + b_i\right) + \beta_i,$$

Learning all blocks

The generator i loss becomes :

$$\mathcal{L}_i = \frac{1}{2}\mathbb{E}_{x\sim p(\mathbf{x_i})}\left[\log(\mathcal{D}(x|\mathbf{x}_{\setminus i}))\right] - \frac{1}{2}\mathbb{E}_{z\sim p(E_i)}\left[\log(1-\mathcal{D}(\mathcal{G}(z|\mathbf{x}_{\setminus i})))\right] \tag{3}$$

Thus the total loss on generators:

$$\mathcal{L} = \underbrace{\sum_{i}^{d}\mathcal{L}_i}_{\text{adversarial generation loss}} + \underbrace{\sum_{i,j,i\neq j}^{d}|a_{ij}|}_{L_1\text{regularization}} \tag{4}$$

Output : adjacency matrix

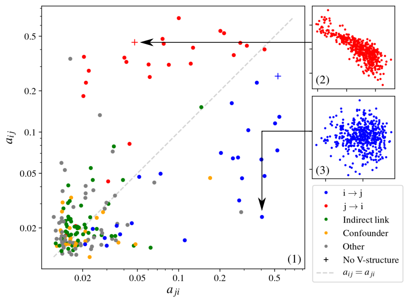
The matrix of a_{ij} represents the adjacency matrix of the graph.

$$\mathbf{A} = \underbrace{\left.\begin{pmatrix} 0 & a_{12} & a_{13} & \cdots & a_{1d} \\ \vdots & \ddots & & & \\ a_{i1} & \cdots & 0 & \cdots & a_{id} \\ \vdots & & & \ddots & \\ a_{d1} & a_{d2} & \cdots & a_{d(d-1)} & 0 \end{pmatrix}\right\} d \ \text{ rows}}_{d \ \text{ columns}}$$

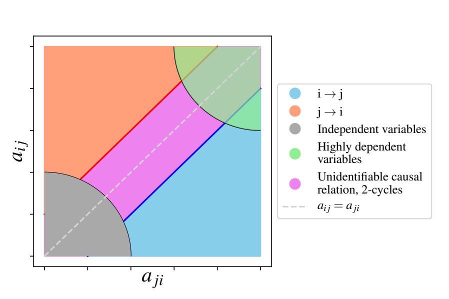
$$\boxed{X_i \to X_j \text{ if } a_{ij} < a_{ji}}$$

Experiments

Interpretation of A



(a) Projections of aij



(b) Causality map

Datasets

1. *Linear*: $X_i = \sum_{j \in \text{Pa}(i)} a_{i,j} X_j + E_i$.
2. *Sigmoid AM*: $X_i = \sum_{j \in \text{Pa}(i)} f_{i,j}(X_j) + E_i$, where $f_{i,j}(x_j) = a \cdot \frac{b \cdot (x_j + c)}{1 + |b \cdot (x_j + c)|}$ with $a \sim \text{Exp}(4) + 1$, $b \sim \mathcal{U}([-2, -0.5] \cup [0.5, 2])$ and $c \sim \mathcal{U}([-2, 2])$.
3. *Sigmoid Mix*: $X_i = f_i(\sum_{j \in \text{Pa}(i)} X_j + E_i)$, where f_i is as in the previous bullet-point.
4. *GP AM*: $X_i = \sum_{j \in \text{Pa}(i)} f_{i,j}(X_j) + E_i$ where $f_{i,j}$ is an univariate Gaussian process with a Gaussian kernel of unit bandwidth.
5. *GP Mix*: $X_i = f_i([X_{\text{Pa}(i)}, E_i])$, where f_i is a multivariate Gaussian process with a Gaussian kernel of unit bandwidth.
6. *Polynomial*: $X_i = \sum_{j \in \text{Pa}(i)} f_{i,j}(X_j) + E_i$, or $X_i = \sum_{j \in \text{Pa}(i)} f_{i,j}(X_j) \cdot E_i$, where $f_{i,j}$ is a random polynomial with a random degree in $[1, 4]$.
7. *NN*: $X_i = f_i(X_{\text{Pa}(i)}, E_i)$, where f_i a random single-hidden-layer neural network with 20 ReLU units.

Artificial graphs

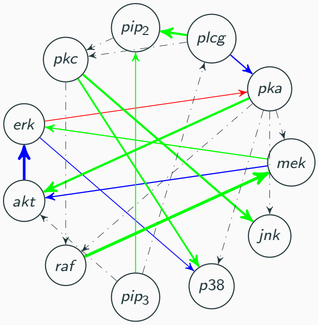
Table 1: Mean Average Precision (AP) on acyclic graphs. AP is between 0 and 1, 1 is best. Bold denotes best performance on the graph size. Underline denotes statistical significance at $p = 10^{-2}$. 55C with the HSIC test has not been evaluated on the graphs of 100 variables, being too computationally expensive.

	PC Gauss		PC HSIC		GES		MMHC		DAGL1		LINGAM		CAM		SAM	
Graph size	20	100	20	100	20	100	20	100	20	100	20	100	20	100	20	100
Linear	0.36	0.39	0.29	/	0.40	0.41	0.36	0.40	0.30	0.28	0.31	0.15	0.29	0.36	**0.49**	**0.46**
Sigmoid AM	0.28	0.24	0.33	/	0.18	0.12	0.31	0.28	0.19	0.15	0.19	0.09	0.72	**0.69**	**0.73**	0.57
Sigmoid Mix	0.22	0.21	0.25	/	0.21	0.15	0.22	0.23	0.16	0.17	0.12	0.05	0.15	0.19	**0.52**	**0.51**
GP AM	0.21	0.17	0.35	/	0.19	0.08	0.21	0.18	0.15	0.12	0.17	0.08	**0.96**	**0.95**	0.74	0.69
GP Mix	0.22	0.17	0.34	/	0.18	0.09	0.22	0.17	0.19	0.12	0.14	0.05	0.61	0.60	**0.66**	**0.66**
Polynomial	0.27	0.27	0.31	/	0.20	0.14	0.11	0.03	0.26	0.29	0.32	0.13	0.47	0.55	**0.65**	**0.56**
NN	0.40	0.40	0.38	/	0.42	0.37	0.11	0.03	0.43	0.50	0.36	0.19	0.22	0.32	**0.60**	**0.55**
Execution time	1s	23s	10h	>50h	<1s	5s	<1s	4s	2s	40s	2s	30s	2.5h	13h	1.2h	4h

Precision/recall score - causal protein network

Table 2: Mean Average Precision (AP) Cyto

	CCD	PC Gauss	GES	MMHC	DAGL1	LINGAM	CAM	CAM
Cyto	0.21	0.16	0.14	0.20	0.22	0.16	0.28	**0.31**



Towards improving SAM

What does represent $a_{i,j}$?

Confusion between:

1. The impact/amplitude of the causal effect
2. The existence of a causal relationship

Example :

$$X = 0.1 * Y + \mathcal{N}_{0,1}$$

Solution: Gumbel Softmax on edges

- By introducing a probability on whether the edge will be present or not, instead of the continous a_{ij}.
- At each epoch a new graph $\widehat{\mathcal{G}}$, where each edge (i, j) is drawn according to Bernoulli distribution with parameter p_{ij}
- These lambdas wil be either 1 or 0 depending on the sampling made on a learned parameter p_{ij} through backpropagation [Jang et al., 2016]

$$a_{i,j} = \text{one hot}(\text{argmax}(g_0 + log(1 - p_{ij}), g_1 + logp_{ij})) \tag{5}$$

A skeleton recovery phase for better scaling (on big datasets)

Like many algorithms, the addition of a pruning phase before the algorithms, improves the consistency, computational efficiency and robustness of the results.

Adding such phase on SAM would help it to scale better (currently d^2)

An (optional) DAG constraint

If was generated by a DAG (Directed Acyclic Graph), an addtitional term can be added to recover a DAG [Zheng et al., 2018]:

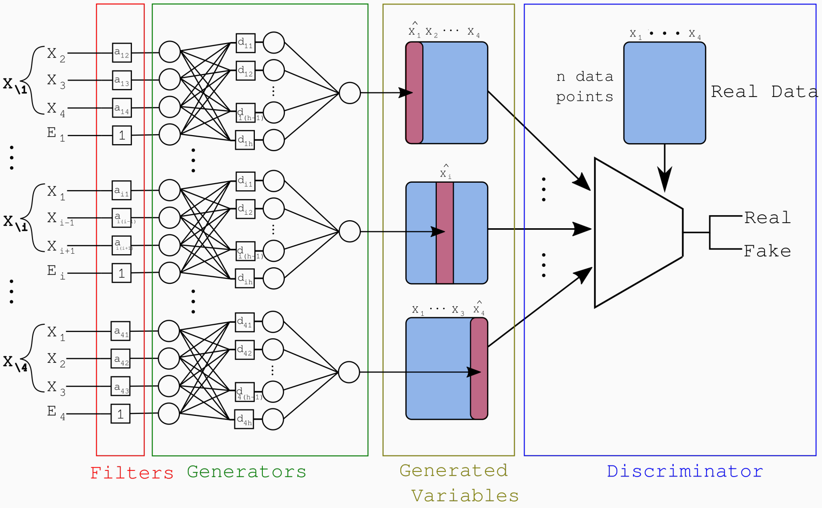
$$\mathcal{L}_{DAG} = \sum_{k=1}^{d} \frac{tr(\mathbf{A}^k)}{k!}$$

An automatic regularization of the functional complexity

Before, the number of hidden units was fixed and dependent of the mecanism complexity that generated the data.

Adding a sampling of units, like the *a* coefficients, adds a control term for this functional complexity.

$$\hat{X}_j = m_j^\top \tanh\left(\bar{W}_j^\top (a_j \odot X) + n_j E_j + b_j\right) \odot d_j + \beta_j, \qquad (6)$$



X\1
X 2
X 3
X 4
E 1
X\i
X 1
X i-1
X i+1
E i
X\4
X 1
X 2
X 3
E 4
Real Data
n data points
Real
Fake
Filters
Generators
Generated Variables
Discriminator

What is usable right now

All the presented framework is available on GitHub at :
https://github.com/Diviyan-Kalainathan/CausalDiscoveryToolbox
It includes multiple algorithms as well as tools for graph structure.

Questions?

Hoyer, P. O., Janzing, D., Mooij, J. M., Peters, J., and Schölkopf, B. (2009).
Nonlinear causal discovery with additive noise models.
In *Advances in neural information processing systems*, pages 689–696.

Jang, E., Gu, S., and Poole, B. (2016).
Categorical reparameterization with gumbel-softmax.
arXiv preprint arXiv:1611.01144.

Spirtes, P., Glymour, C. N., and Scheines, R. (2000).
Causation, prediction, and search.
MIT press.

Tsamardinos, I., Brown, L. E., and Aliferis, C. F. (2006).
The max-min hill-climbing bayesian network structure learning algorithm.
Machine learning, 65(1):31–78.

Zheng, X., Aragam, B., Ravikumar, P., and Xing, E. P. (2018).
DAGs with NO TEARS: Smooth Optimization for Structure Learning.
ArXiv e-prints.