

Causal Modeling

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



TAU, Université Paris Saclay, France

Jan. 2020

Motivations

Artificial Intelligence / Machine Learning

A Case of Irrational Scientific Exuberance

- | Underspecified goals Big Data cures everything
- | Underspecified limitations Big Data can do anything (if big enough)
- | Underspecified caveats Big Data and Big Brother

Wanted: An AI with common decency

- | Fair no biases
- | Accountable models can be explained
- | Transparent decisions can be explained
- | Robust w.r.t. malicious examples

ML & AI, 2

In practice

- | Data are ridden with biases
- | Learned models are biased (prejudices are transmissible to AI agents)
- | Issues with robustness
- | Models are used out of their scope

More

- | C. O'Neill, *Weapons of Math Destruction*, 2016
- | Zeynep Tufekci, *We're building a dystopia just to make people click on ads*, Ted Talks, Oct 2017.

ML yields discriminative or generative modelling

Given a training set

iid samples $P(X; Y)$

$$E = f(\mathbf{x}_i; y_i); \mathbf{x}_i \in \mathbb{R}^d; i \in \{1; n\}$$

Find

- | Supervised learning: $\hat{h} : X \rightarrow Y$ or $\hat{p}(Y|X)$
- | Generative model $\hat{p}(X; Y)$

Predictive modelling might be based on correlations

If umbrellas in the street, Then it rains



The implicit big data promise:

If you can predict what will happen,
then how to make it happen what you want ?

Knowledge ! Prediction ! Control

ML models will be expected to support interventions:

Intervention $do(X = a)$ forces variable X to value a

- | health and nutrition
- | education
- | economics/management
- | climate

The implicit big data promise, 2

Intervention

Pearl 2009

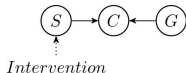
Direct cause $X \rightarrow Y$

$$P_{Y \text{ do}(X=a; \mathbf{Z}=c)} \neq P_{Y \text{ do}(X=b; \mathbf{Z}=c)}$$

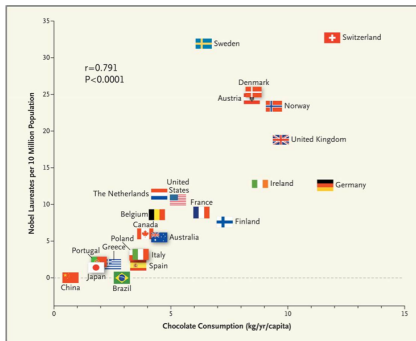
Example

C: Cancer, S : Smoking, G : Genetic factors

$$P(C \text{ do } S = 0; G = 0g) \neq P(C \text{ do } S = 1; G = 0g)$$



Correlations do not support interventions



F. H. Messerli: *Chocolate Consumption, Cognitive Function, and Nobel Laureates*, N Engl J Med 2012

Causal models are needed to support interventions

*Consumption of chocolate enables to predict # of Nobel prizes
but eating more chocolates does not increase # of Nobel prizes*

Predictive model δ Causal model

Consider

$$\begin{array}{ll} X; E_Y; E_Z & \text{Uniform}(0;1); \\ Y & 0.5X + E_Y; \\ Z & Y + E_Z; \end{array}$$

with $E_Y; E_Z \sim N(0;1)$ (noise)

Predicting Y

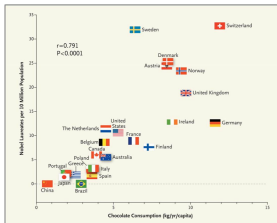
$$\hat{Y} = 0.25X + 0.5Z$$

If interpreted as a causal model, suggests that Y depends on Z .

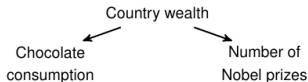
Issue

Causes can often be predicted from their effects

When correlations do not imply causality



F. H. Messeri: Chocolate Consumption, Cognitive Function, and Nobel Laureates. N Engl J Med 2012



Tentative explanation: confounders

- | Both effects of a same cause, $C \text{ \& } N$.
- | But C and N are conditionally independent given W

$$C \text{ ? } N|W$$

Causality and paradoxes

Facts

- | If mother smokes, child weight tends to be small
- | Tiny child, more health problems
- | However, tiny child AND mother smokes > tiny child

Interpretation mother smoking beneficial to child's health ?

Explaining away

Many possible causes for small child weight

Many of these severely affect child's health (genetic diseases)

Compared to these, mother smoking is rather a good news...

An AI with common decency

Desired properties

- | Fair no biases
- | Accountable models can be explained
- | Transparent decisions can be explained
- | Robust w.r.t. malicious examples

Relevance of Causal Modeling

- | Decreased sensitivity wrt data distribution
- | Support interventions clamping variable value
- | Hopes of explanations / bias detection

Causal Discovery

HOW

- | Gold Standard: perform randomized controlled experiments
- | But these experiments are often costly, unethical or unfeasible
- | Our setting: observational causal discovery
From data, infer causal model.

WHAT FOR

- | Understandable, interpretable, more robust models
- | Prioritize confirmatory experiments: enabling some control
- | Generate new data: privacy and domain-compliant, e.g. for medical training

Motivating applications

Human resources

1. Autonomy / Satisfaction / Productivity
2. Quality of life at work / Economic profitability of firms

Joint project with 'La Fabrique de l'industrie'

Kalainathan et al. 18

Health and Life habits

1. Diet / Diabetes type 2.

Joint project Nutriperso with INRA



State of the art

Causal Modelling

The Causal Discovery Setting

Assume random variables

X_1, \dots, X_d : random variables

and a sample of their joint distribution

$$D = \{ \mathbf{x}_i; i = 1 \dots n \}$$

to be given.

Formal background: Overview

1. Key concepts
2. Framework
3. Approaches

Key concepts: 1. Dependence among pairs of variables

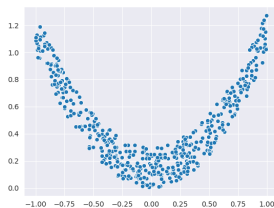
Independent variables X and Y ($X \perp\!\!\!\perp Y$)

$$X \perp\!\!\!\perp Y \text{ iff } P(X; Y) = P(X) \cdot P(Y)$$

Dependency tests

| Correlation

limited to linear dependencies



$$Y = X^2 + E$$
$$\text{Correlation}(X; Y) = 0$$

Key concepts: 1. Dependence among pairs of variables

Independent variables X and Y ($X \perp\!\!\!\perp Y$)

$$X \perp\!\!\!\perp Y \text{ iff } P(X; Y) = P(X)P(Y)$$

Dependency tests

- | Correlation limited to linear dependencies
- | HSIC, Hilbert-Schmitt Independence Criterion

Gretton et al. 05

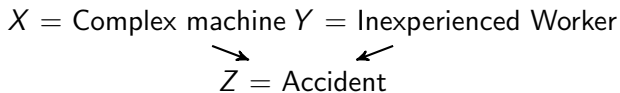
$$HSIC(P_{XY}; F; G) := \|C_{XY}\|_{HS}^2$$

where $\| \cdot \|_{HS}$ denotes the Hilbert-Schmidt norm, and C_{XY} a kernel based covariance operator and $F; G$ two RKHSs.

Key concepts: 2. Conditional Dependence/Independence

Conditional independence a.k.a. hidden confounder

Conditional dependence a.k.a. V-structure



X and Y are independent; but given $Z = \textit{true}$ they are not independent (either the machine is complex or the worker is inexperienced...)

Definition of causal relationship

Definition of intervention

$do(X = 1)$ forces variable X to value 1

Pearl 09

Definition of causal relationship

X is a direct cause of Y ($X \rightarrow Y$) iff
all other variables Z being constant,

$$P_{Y|do(X=1; \dots; Z=c)} \neq P_{Y|do(X=0; \dots; Z=c)}$$

Definition of causal relationship

Definition of intervention

$do(X = 1)$ forces variable X to value 1

Pearl 09

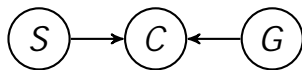
Definition of causal relationship

X is a direct cause of Y ($X \rightarrow Y$) iff
all other variables Z being constant,

$$P_{Y|do(X=1; \dots; Z=c)} \neq P_{Y|do(X=0; \dots; Z=c)}$$

Example C : Cancer, S : Smoking, G : Genetic factors.

$$P(C|do(S = 0; G)) \neq P(C|do(S = 1; G))$$



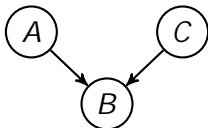
Intervention

Markov equivalence class and V-structure

Markov Equivalent Class: $A \perp\!\!\!\perp C \mid B$ and $A \perp\!\!\!\perp C$



V-Structure: $A \perp\!\!\!\perp C \mid B$ and $A \perp\!\!\!\perp C$

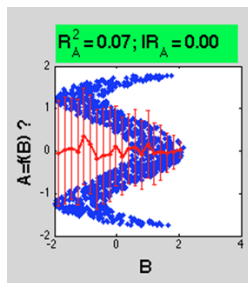
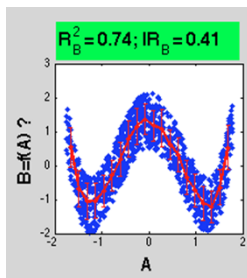


Key concepts: 3. Causality with distributional asymmetry

Leveraging Occam's razor principle;

Janzig 19

! the causal model as the one being the simplest model that fits the data.



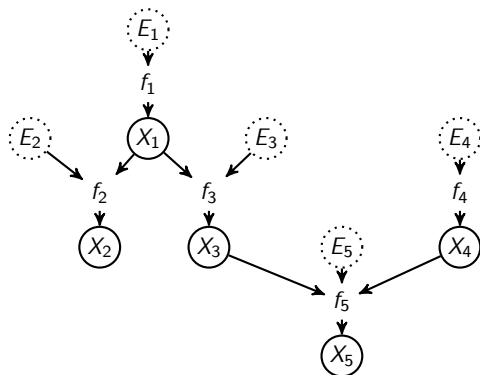
Framework: Functional Causal Models (FCMs)

Given $X_1; \dots; X_d$,

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

with $X_{\text{Pa}(i;G)}$ the set of parents of X_i in G (= causes of X_i),
 E_i a random independent noise variable modeling the unobserved
other causes,

f_i a deterministic function: the causal mechanism



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

Functional Causal Models, 2

Markov decomposition

$$P(X_1; \dots; X_d) = \prod P(X_i | X_{\text{Pa}(i; G)})$$

Usual Assumptions

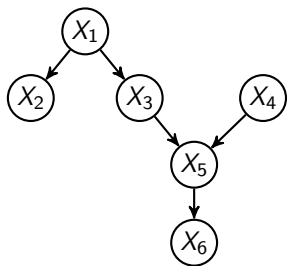
Causal Sufficiency: no unobserved confounders

Causal Markov: all d -separations in the causal graph G imply conditional independences in the observational distribution P

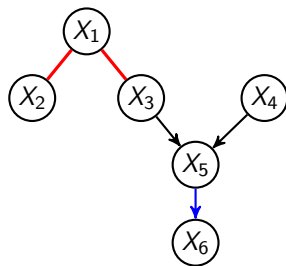
Causal Faithfulness: all conditional independences in P imply d -separations in G .

Key approach 1: Constraint-based methods

Constraint-based methods, through V-Structures and constraint propagation, output a **CPDAG** (Completed Partially Directed Acyclic Graph).



(a) The exact DAG of G .



(b) The CPDAG of G .

Ex: Peter-Clark Algorithm (PC)

Non-linear extensions (CI tests): PC-HSIC (KCI-test), PC-RCIT

Spirtes et al. 00

Zhang 12, Strobl 17

Key approach 2: Score-based methods

Objective function to optimize such as the Bayesian Information Criterion (BIC):

$$BIC(G) = -2 \ln L + k \ln n$$

with L : Likelihood of the model, k : number of parameters, n :
Number of samples

The graph is optimized with the operators:

- | add edge
- | remove edge
- | revert edge

Ex: Greedy Equivalence Search (GES)

Limitations

- | Computational cost dependent on the type of test/scoring method used
- | Data hungry
- | Identifiability issues

Example

$$\begin{array}{ll} X_1; E_{X_1}; E_{X_2} & \text{Uniform}(0;1); X_1 \perp\!\!\!\perp E_{X_1}; Y \perp\!\!\!\perp E_{X_2} \\ Y & 0.5X_1 + E_{X_1}; \\ X_2 & Y + E_{X_2}; \end{array}$$



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

Key approach 3: Global optimization

Assuming linear causal mechanisms, the causal mechanisms can be formulated in terms of linear algebra.

$$\mathbf{X} = B^T \mathbf{X} + E$$

And estimate the B matrix, through ICA for LiNGAM

Shimizu 06, Hyvarinen 99

! Graphical models

Pearl 09, Friedman 08

Ex: Max-Min Hill-Climbing (MMHC)

Tsamardinos 06

Concave penalized Coordinate Descent (CCDr)

Aragam 15

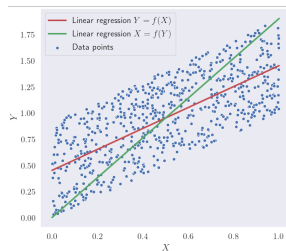
Key approach 4: Exploiting asymmetries in the distribution

! If no v-structure available or causal discovery with 2 variables:
leverage asymmetries in the distributions.

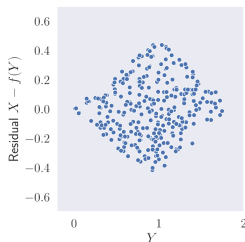
Additive noise model (ANM):

Hoyer 09

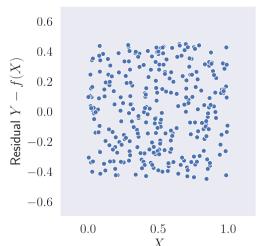
$$Y = f(X) + E$$



Original data



Residuals of $X=g(Y)$



Residuals of $Y=f(X)$

Ex: Post Non-Linear model (PNL), GPI

Zhang 10, Stegle 10

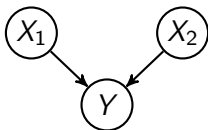
Limitations of asymmetry-based approaches

- | Restrictive assumptions on the type of causal mechanisms
- | Does not take into account conditional independence relations.

Zhang 09

Example

$$\begin{array}{l} X_1; X_2; E_{X_1} \\ Y \end{array} \quad \text{Gaussian}(0; 1); X_1 \perp\!\!\!\perp E_{X_1}; X_2 \perp\!\!\!\perp E_{X_1}$$
$$Y = 0.5X_1 + X_2 + E_{X_1}$$



$(X_1; Y)$ 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

Key approach 5: A machine learning-based approach

Guyon et al, 2014-2015

Pair Cause-Effect Challenges

- | Gather data: a sample is a pair of variables ($A_i; B_i$)
- | Its label τ_i is the “true” causal relation (e.g., age “causes” salary)

Input

$$E = f(A_i; B_i; \tau_i); \tau_i \in \{!; ?\}$$

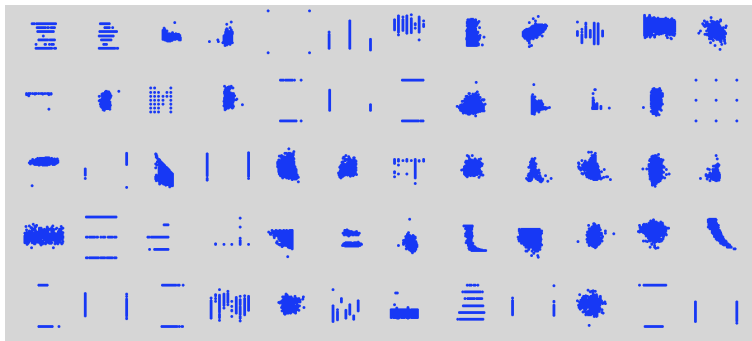
Example $A_i; B_i$	Label τ_i
A_i causes B_i	!
B_i causes A_i	!
A_i and B_i are independent	?

Output

using supervised Machine Learning

Hypothesis : $(A; B) \mapsto$ Label

Key approach 5: A machine learning-based approach, 2



The Cause-Effect Pair Challenge

Learn a causality classifier (causation estimation)

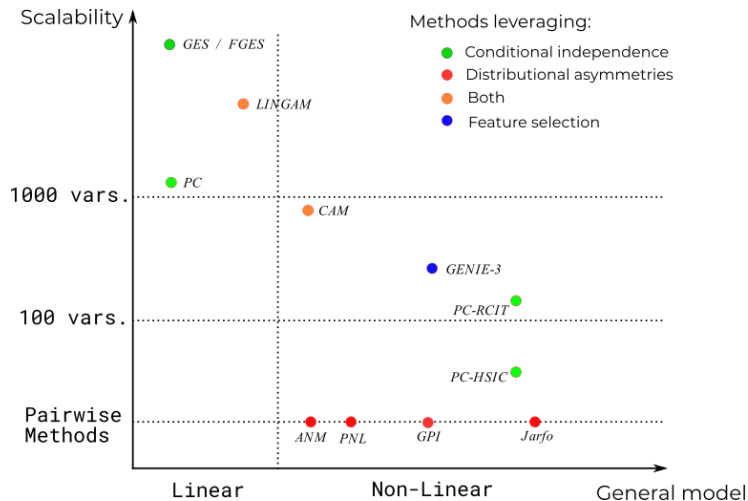
- Like for any supervised ML problem from images ImageNet 2012



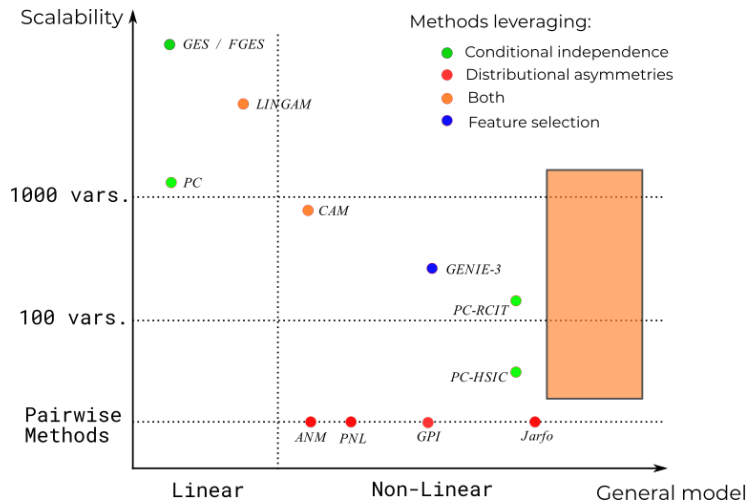
More

- Guyon et al., eds, *Cause Effect Pairs in Machine Learning*, 2019.

State of the art: summary



State of the art: summary



Causal Generative Neural Networks

Causal Generative Neural Networks (CGNN): Overview

Assumptions:

- | Input: Graph skeleton with L edges
- | Continuous data: $X_1 \dots, X_d$ real valued

Problem posed:

- | Combinatorial optimization problem of dimension L
- | For each candidate in $\mathcal{F} = \{1, \dots, 2^L\}$, find each causal mechanism

Approach:

- | Causal mechanisms f_i approximated as a neural net.
- | Loss function: Maximum Mean Discrepancy (MMD) (distance original vs generated data);
- | Hyperparameter: number n_h of neurons in f_i

Modeling FCMs with generative neural networks

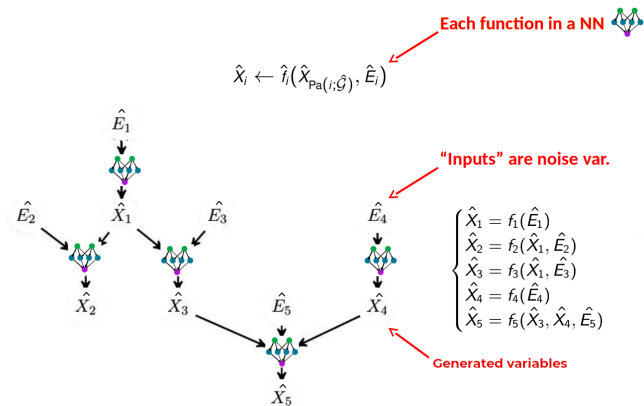
- Idea: approximate the continuous mechanisms $f_1; \dots; f_d$ with a set of one hidden layer neural networks $\hat{f} = (\hat{f}_1; \dots; \hat{f}_d)$

Modeling FCMs with generative neural networks

- | Idea: approximate the continuous mechanisms $f_1; \dots; f_d$ with a set of one hidden layer neural networks $\hat{f} = (\hat{f}_1; \dots; \hat{f}_d)$
- | Estimate FCMs C as $\hat{C} = (\hat{G}; \hat{f})$:

$$\hat{X}_i \sim \hat{f}_i(\hat{X}_{\text{Pa}(i; \hat{G})}; E_i); E_i \sim N(0; 1) \quad (1)$$

Generative neural networks as a FCM



For each candidate $(\hat{G}; \hat{f})$, generate samples \hat{X} ;
Loss = difference between original distribution, generated distribution

Learning Metric: Maximum Mean Discrepancy (MMD)

Kernel-based loss evaluating a "distance" between empirical distributions:

Gretton 05

- | Generated data $\hat{\mathbf{X}} = \hat{\mathbf{x}}_i; i = 1 \dots n^0$
- | True data $\mathbf{X} = \mathbf{x}_i; i = 1 \dots n$

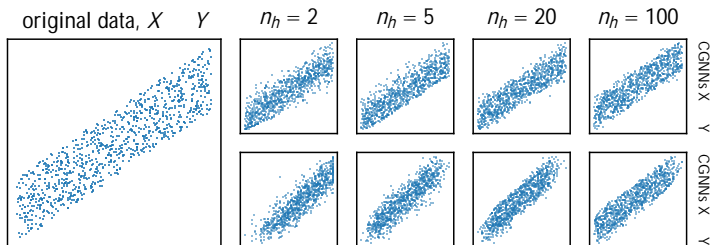
$$MMD(\hat{\mathbf{X}}; \mathbf{X}) = \frac{1}{n^2} \sum_{i,j} k(\mathbf{x}_i; \mathbf{x}_j) + \frac{1}{n^0} \sum_{i,j} k(\hat{\mathbf{x}}_i; \hat{\mathbf{x}}_j) - \frac{2}{nn^0} \sum_{i,j} k(\hat{\mathbf{x}}_i; \mathbf{x}_j)$$

with $k(\mathbf{u}; \mathbf{v}) = \exp\left(-\frac{\gamma}{2} \|\mathbf{u} - \mathbf{v}\|^2\right); \gamma \in [2 \cdot 10^{-2}; \dots; 10^2]$

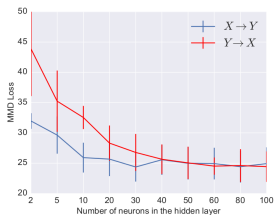
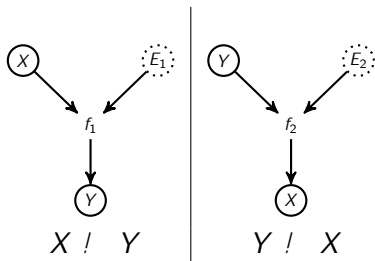
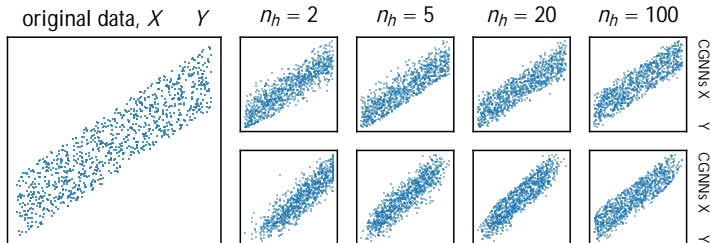
A linear approximation \hat{MMD} leveraging random projections has been proposed

Lopez-Paz et al. 16

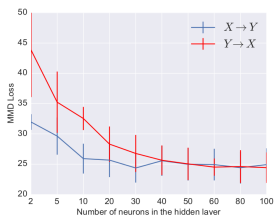
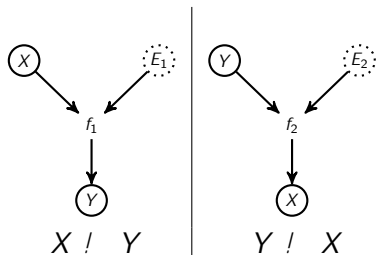
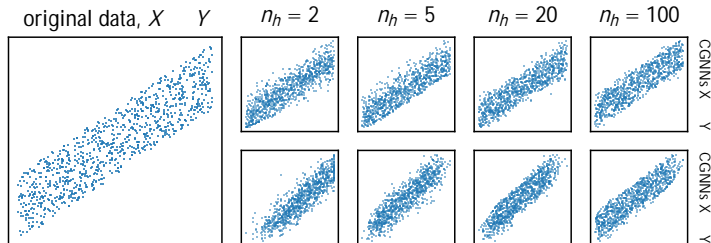
Adjusting number of hidden units n_h



Adjusting number of hidden units n_h



Adjusting number of hidden units n_h



) Causal direction not identifiable if n_h too high

General algorithm

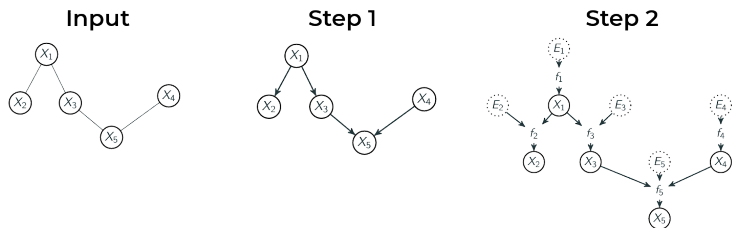
Input = Continuous Data + Graph skeleton

1. Init: Pairwise orientation + DAG recovery (remove cycles heuristic)
2. Iteratively until the stopping criterion is met:
 - | Reverse an edge at random that does not create a cycle
 - | Retrain CGNN using backpropagation
 - | If the resulting MMD loss is better, replace the current best solution

General algorithm

Input = Continuous Data + Graph skeleton

1. Init: Pairwise orientation + DAG recovery (remove cycles heuristic)
2. Iteratively until the stopping criterion is met:
 - | Reverse an edge at random that does not create a cycle
 - | Retrain CGNN using backpropagation
 - | If the resulting MMD loss is better, replace the current best solution



Experimental setting

- | Benchmarks:
 - | Simulated data: $X_i = f_i(X_{Pa(i;G)}; E_i); \delta_i \in [1; d]$,
with f_i : Polynomials, Gaussian processes with additive and multiplicative noise
 - | Biological data : SynTReN Gene expression, Real protein network

Sachs 05

- | All methods are given the true skeleton
- | Performance indicator: Area under the Precision Recall Curve (number of identified edges)

Experimental setting

- | Benchmarks:

- | Simulated data: $X_i = f_i(X_{Pa(i;G)}; E_i); \delta_i \in [1; d]$,
with f_i : Polynomials, Gaussian processes with additive and multiplicative noise

- | Biological data : SynTREN Gene expression, Real protein network

Sachs 05

- | All methods are given the true skeleton

- | Performance indicator: Area under the Precision Recall Curve
(number of identified edges)

- | Baselines:

- | PC, PC-HSIC (KCI-test)

Spirtes 00, Zhang 11

- | ANM

Hoyer 09

- | Jarfo

Fonollosa 16

- | GES

Chickering 02

- | LiNGAM

Shimizu 06

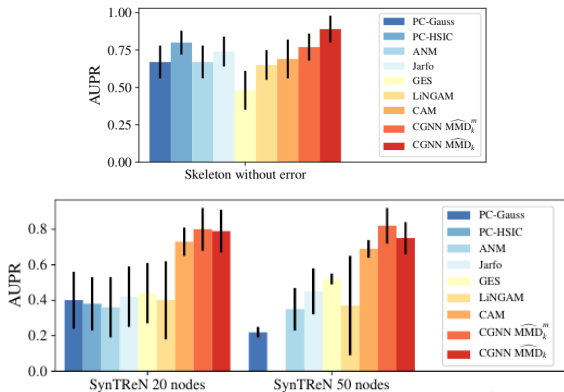
- | CAM

Buhlman 14

- | CGNN: $n_h \in [5; 20]; epochs = 2000; \eta = 0.01$

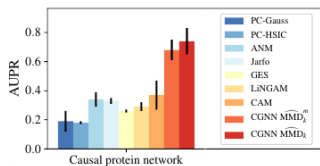
- | $\hat{MMD}_k^m; m = 300$ (Linear approx of MMD)

Experimental validation: Generated datasets

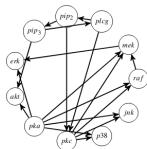
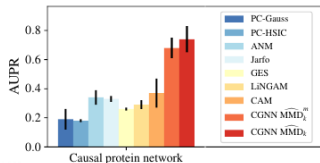


All methods are given the true skeleton.

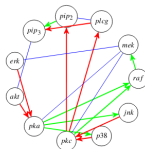
Experimental validation: Real data



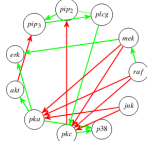
Experimental validation: Real data



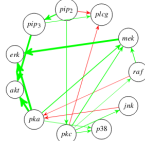
(a) Ground truth



(b) GES



(c) CAM



(d) CGNN

Color: green: ok ; red: wrong; blue: unknown, Edge width: confidence

CGNN

PROS:

- | **[UNIVERSALITY]** power of NN (universal approximators)
- | **[UNIFICATION]** unification of causal discovery principles (CI and DA)

CONS:

- | **[SKELETON KNOWLEDGE NEEDED]** the method requires the initial knowledge of the graph skeleton (though edge orientation is robust against skeleton mistakes)
- | **[COMPUTATIONAL COST]** the method is computationally costly (30h for 50 variables) which in practice required us to perform sub-optimal greedy optimizations
- | **[SENSITIVITY]** the method is sensitive to hyper-parameter selection (including number of neurons)

Structural Agnostic Modeling

Structural Agnostic Model (SAM): Overview

Assumptions:

- | Continuous data
- | Causal sufficiency (no hidden confounder)

Goal:

- | Learn end-to-end the graph structure and the causal mechanisms

Approach:

- | A global loss
- | accounting for structural and functional complexity
- | accounting for model fitness through an adversarial mechanism

Finding the causes for each variable

$$X_j = f_j(X_{-j}; E_j); \quad (2)$$

Finding the causes for each variable

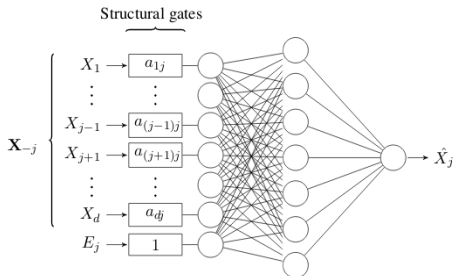
$$X_j = f_j(X_{-j}; E_j); \quad (2)$$

Goal: Find the causes = a sparse network it generates

Finding the causes for each variable

$$X_j = f_j(X_{-j}; E_j); \quad (2)$$

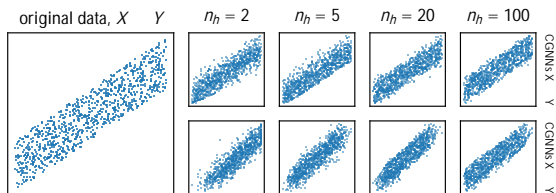
Goal: Find the causes = a sparse network it generates



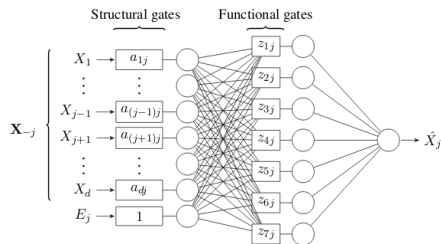
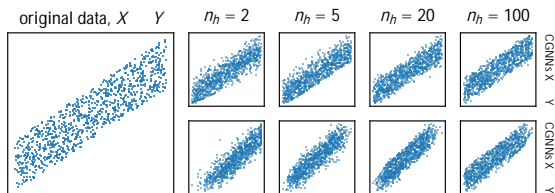
! Enforcing sparsity through L_0 penalization

Leray 99, Maddison 16, Jang 16

Regularization of the complexity of the mechanisms



Regularization of the complexity of the mechanisms



! Enforcing the sparsity of the mechanisms through L_0 penalization

General architecture and loss of SAM

! Adversarial loss

goodfellow2014generative

General architecture and loss of SAM

! Adversarial loss

goodfellow2014generative

Loss of SAM

Learning criterion to minimize:

$$S(\theta; \hat{f}; D) = \underbrace{E_x \sum_{h=1}^H \log q(x; \theta)}_{\substack{\text{Log likelihood} \\ \text{estimated by the discriminator}}} + \underbrace{\lambda \sum_{k=1}^K \sum_{z=1}^Z z}_{\text{Regularization}} \quad (3)$$

where

- | $kAK_1 = \sum_{i,j=1..d} a_{i,j}$: total number of **edges** in \mathcal{G}
! Structural complexity.
- | $kZK_1 = \sum_{j=1,\dots,d} \sum_{h=1,\dots,n_h} Z_{j,h}$: total number of **active units** in \mathcal{P}
! Functional complexity.

Final learning objective

$$\begin{aligned}
 S(\theta; \hat{f}; D) = & \sum_{j=1}^{\mathcal{X}^d} I(X_j; X_{\overline{\text{Pa}(j)}; \theta} | X_{\text{Pa}(j); \theta}) + \sum_{A \in \mathcal{A}} \|A\|_1 \\
 & \underbrace{\hspace{10em}}_{\text{Structural score}} \\
 + & \sum_{j=1}^{\mathcal{X}^d} D_{\text{KL}}[p(x_j | x_{\text{Pa}(j); \theta}) \| q(x_j | x_{\text{Pa}(j); \theta}; j)] + \sum_{Z \in \mathcal{Z}} \|Z\|_1 \\
 & \underbrace{\hspace{10em}}_{\text{Functional score}} \\
 + & \sum_{k=1}^{\mathcal{D}} \frac{\text{tr } A^k}{k!} \\
 & \underbrace{\hspace{10em}}_{\text{Acyclicity constraint}}
 \end{aligned}$$

Zheng 18

with I the mutual information and D_{KL} the Kullback-Leibler divergence

Properties of the score

Theorem 1: Identification to the Markov Equivalence Class

Under Causal Markov and faithfulness assumptions, the DAG minimizing the **structural score** belongs to the Markov equivalence class of the true graph G (CPDAG of G)

Properties of the score

Theorem 1: Identification to the Markov Equivalence Class

Under Causal Markov and faithfulness assumptions, the DAG minimizing the **structural score** belongs to the Markov equivalence class of the true graph G (CPDAG of G)

Theorem 2: Identification of the DAG

Under additional assumptions, the DAG minimizing also **the functional score** is exactly the DAG

Experimental setting

- | Benchmarks:

- | Simulated data (20 and 100 Variables):

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

f_i : Linear, Gaussian processes with additive (GP AM) and multiplicative noise (GP Mix), Sigmoid functions (Sigmoid AM/Sigmoid Mix), Neural networks with randomized weights (NN).

- | Biological data : SynTReN Gene expression , Real protein network

Sachs 05

- | Performance indicator: Area under the Precision Recall Curve

Experimental setting

| Benchmarks:

| Simulated data (20 and 100 Variables):

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

f_i : Linear, Gaussian processes with additive (GP AM) and multiplicative noise (GP Mix), Sigmoid functions (Sigmoid AM/Sigmoid Mix), Neural networks with randomized weights (NN).

| Biological data : SynTReN Gene expression , Real protein network

Sachs 05

| Performance indicator: Area under the Precision Recall Curve

| Baselines:

| PC, PC-HSIC (KCI-test)

Spirites 00, Zhang 11

| PC-RCIT/RCOT

Strobl 17

| ANM

Hoyer 09

| Jarfo

Fonollosa 16

| GES

Chickering 02

| LiNGAM

Shimizu 06

| CAM

Buhlman 14

| MMHC

Tsamardinos 06

| CCDr

Aragam 17

| GENIE2

Irrthum 10

Experimental setting (2)

- | Hyperparameters of SAM:
 - | $\lambda_r = 0:01$
 - | $\lambda_A = 0:01$
 - | $\lambda_z = 10^{-5}$
- | Lesion study (impact of neural vs linear mechanisms and mean square error vs adversarial loss):
 - | SAM-mse-linear: Linear mechanisms and a MSE loss
 - | SAM-linear: Linear mechanisms and a GAN Setting
 - | SAM-mse: Non-linear mechanisms and a MSE Loss

Experimental results: Generated datasets (20 variables)

CAM is especially tailored for Gaussian processes with additive noise;
and GES for linear mechanisms

Experimental results: Generated datasets (100 variable

Results on biological data

Syntren Dataset

Sachs dataset

Ablation studies

Both the non-linear mechanisms and the adversarial network are required to attain maximum performance

Computational time (graph of 100 variables)

AP	Time in s. (CPU)	Time in s. (GPU)
PC-Gauss	13	
PC-HSIC	-	
PC-RCOT	31 320	
PC-RCIT	46 440	
GES	1	
GIES	5	
MMHC	5	
LiNGAM	5	
CAM	45 899	
CCDr	3	
GENIE3	511	
SAM-lin-mse	3 076	74
SAM-mse	18 180	118
SAM-lin	24 844	1 980
SAM	24 844	2 041

Applications

Applications: 1. Human Resources

Causal Modeling and Human Resources

Known:

- A Quality of life at work employee's perspective
- B Economic performance firm's perspective
- | ... are correlated

Question: Are there causal relationships ?

A ! B ; or B ! A; or $C = C ! A$ and C ! B

Data

- | Polls from Ministry of Labor
- | Gathered by Group Alpha Seca (trade union advisor)
- | Tax files + social audits for 408 firms

Economic sectors: low tech, medium-low, medium-high and

Variables

Economic indicators

- | Total number of employees
- | Capitalistic intensity, Total payroll, Gini index
- | Average salary (of workers, technicians, managers)
- | Productivity, Operating profits, Investment rate

People

- | Average age, Average seniority, Physical effort,
- | Permanent contract rate, Manager rate, Fixed-term contract rate, Temporary job rate, Shift and night work, Turn-over
- | Vocational education effort, duration of stints, Average stint rate (for workers, technicians, managers);

Variables, cont'd

Quality of life at work

- | Frequency & Gravity of work injuries, Safety expenses, Safety training expenses
- | Absenteism (diseases), Occupational-related diseases
- | Resignation rate, Termination rate, Participation rate
- | Subsidy to the works council

Men/Women

- | Percentage of women (employees, managers)
- | Wage gap between women and men (average, for workers, technicians, managers)

General Causal Relations

Access to training %

- | & Gravity of work injuries
- | & Occupational-related diseases

Termination rate %

- | % Absenteism (diseases)

Percentage of managers %

- | % Access to training
- | & Shift or night working hours

Age %

Global relations between QLW and performance ?

Failure

- | Nothing conclusive

Interpretation

- | Exist confounders (controlling QLW and performance)!
- | A and C ! B
- | One such confounder is the activity sector
- | In different activity sectors, causal relations are different (hampering their identification)
- |) Condition on confounders

Low-tech sector

- | Resignation rate%, Productivity &
- | Average salary%, Productivity % very significant
- | Occupational-related diseases%, Productivity &
- | Temporary job rate%, Gravity of work injuries%
- | Permanent contract rate%, Safety training&
- | Duration training stints%, Termination rate&

Outcomes & Limitations

Causal modeling and exploratory analysis

- | Efficient Iterating of plausible relations (several orders of magnitude);
- | Complementary w.r.t. visual inspection (experts can be fooled and make sense of correlations & hazards);
- | Multi-factorial relations ? yes; but even harder to interpret.

Not a ready-made analysis

- | Causal relations must be
 - | interpreted
 - | confirmed by field experiments; polls; interviews.

Applications: 2. Food and Health

A data-driven approach to individual dietary recommendations

Context

- | Long-term goal: Personalized dietary recommendations
- | Requirement: identify risk index associated to food products
- | At a coarse-grained level (lipid, protein, glucid), nothing to see
- | At a ne-grained level: 300+ types of pizzas, ranging from ok to very bad.

The wealth of Kantar data

- | 22,000 households 10 years (this study: 2014)
- | 19M total purchases/year (180,000 products)
- | Socio-demographic attributes, varying size

Beware: data rarely collected as should be...

Raw description can hardly be used for meaningful analysis

- | 170,000 products for 22,000 households
- | Data gathered with (among others) marketing goals where bought, which conditioning
- | Most products are sold by 1 vendor
- | Most families are going to one vendor

Manual pre-processing

- | Consider 10 categories of interest, e.g. bio/non-bio; alcohol yes/no; fresh/frozen
- | Merge products with same categories
- | 170,000! 4,000 products

Example: for beer, we only selected as features of interest: colour (blonde, black, etc.); has-alcohol (yes, no); organic (yes, no)

Methodology

Dimensionality reduction

1. Borrowing Natural Language Processing tools, with
vector of purchase document
food product word
2. Using Latent Dirichlet Association to extract "dietary topics"
Blei et al. 03

Some topics can be directly interpreted The darker the region, the more present the topic (NB: regions are not used to build topics)

Focus: impact of topics on BMI

Left: Bio/organic topic

Top row: Women

Right: Frozen food topic

Bottom row: Men

High weight of Bio topic is correlated with lower BMI ($< 5\%$)
(particularly so for women)

Proposed Methodology

Taking inspiration from Abadie Imbens 06

Target population: "Bio" people = top quantile coordinate on bio topic.

RCT would require a control population

Building a control population finding matches

- | For each bio person, take her consumption (basket of products)
- | Create a falsified consumption⁰ (replacing each bio product with same, but non-bio, product)
- | Find true consumption z^* nearest to z^0 (in LDA space)
- | Let the true person with consumption z^* be called "falsified bio"

Compare bio and "falsified bio" populations wrt BMI

Bio vs Falsi ed Bio populations

Left

- | Projection on the Bio topic (in log scale)
- | (Falsi ed bio population not 0: the bio topic contains e.g. sheep yogurt).

Right

- | BMI Histograms of both bio and falsi ed bio populations
- | Statistically significant difference

Next

Chasing confounders

- | Discriminating bio from "falsified bio" populations w.r.t. socio-professional features: accuracy 60%
- | Candidate confounder: mother education level (on-going study)

Next steps

- | Confirm conjectures using longitudinal data (2015-2016)
- | Interact with nutritionists / sociologists
- | Extend the study to consider the impact of, e.g.
 - | Price of the food
 - | Amount of trans fats
 - | Amount of added sugar

Discussion

Perspectives: Causality analysis and Big Data

Finding the needle in the haystack

- | Redundant variables (e.g. in economics) / un-interesting relations
- | Variable selection
- | Feature construction dimensionality reduction

Beyond causal sufficiency

- | Confounders are all over the place (and many are plausible, e.g. age and size of firm; company ownership and shareholdings)
- | When prior knowledge available, condition on confounders
- | Use causal relationships on latent variables Wang and Blei, 19
to filter causal relationships on initial variables

A python package for observational causal discovery

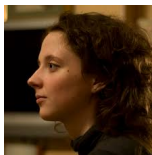
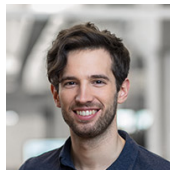
All the presented framework is available on GitHub at :

<https://github.com/Diviyan-Kalainathan/CausalDiscoveryToolbox>

It includes multiple algorithms as well as tools for graph structure.

Accepted at JMLR - Open Source Software

Kalainathan Goudet 19



Thanks to Isabelle Guyon, Diviyen Kalainathan, Olivier Goudet,
David Lopez-Paz,
Philippe Caillou, Paola Tubaro