Stability Analysis of the Structural Agnostic Modeling Method

Ramon Daniel REGUEIRO ESPIÑO

ENS Paris-Saclay

TAU, Inria-Saclay, LISN

Tutor: BARDENET Rémi

Supervisor: LEITE Alessandro
Supervisor 2: POINSOT Audrey

Overview

Stability

Stability: a small perturbation on the inputs of an algorithm does not change too much its output.

Unstable algorithms limitations:

- How can we rely on their results?
- Lack of replicability

Why causality?

- Spurious correlations might lead to wrong interventions.

Figure 1: Smoking may lead to yellow fingers (image generated by DALL-E mini).

Main idea:

- Causality gives more information than correlations

Randomized Controlled Trial (RCT)

RCT:

- Experiment controlled by the researcher.
- The gold standard.

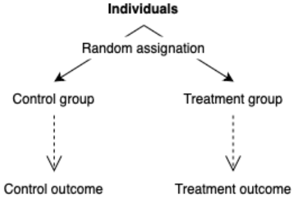

Figure 2: Example of RCT schema.

Not always feasible: e.g. economical or ethical limitations.

Observational data: data without any manipulation.

Internship aim

Structural Agnostic Model (SAM) Kalainathan et al. 2022 model: Generative Adversarial Network (GAN) Goodfellow et al. 2014 model aiming to discover causal relationships.

Main problem:

- Even with the same initialization, SAM is not stable.

Our goal:

- Explore SAM instability.

Main question:

- Under which conditions SAM is stable?

Overview

Functional Causal Model (FCM) Pearl 2003

Functional Causal Model (FCM) Pearl 2003

Considering a set of random variables $X_1, \cdots, X_n$, a FCM is a set of equations

$$x_i = f_i(x_{\text{pa}_i}, u_i) \quad i = 1, \cdots, n. \tag{1}$$

- X_{pa_i}: set of observed variables that directly determine the value of X_i.
- U_i: random variable modelling the noise.
- f_i: causal mechanism.

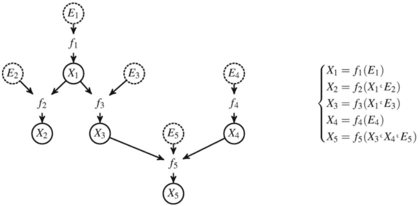

Figure 3: Example of FCM (Figure from Goudet et al. 2018).

Causal graphical models

Causal graph:

- Variables as nodes.
- Each direct edge ($\rightarrow$) is a cause-effect oriented relation.

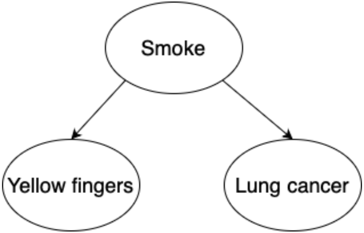

Figure 4: Causal graph relating smoking, yellow fingers and lung cancer.

Remark:

- A causal graph and a Bayesian Network are not equivalent!

Causal Markov Condition (CMC)

Acyclicity: the graph does not have any (direct) cycle.

Directed Acyclic Graph (DAG)

A direct graph without any direct cycle is called a DAG.

Causal Markov Assumption (CMA): For a given causal graph, all the considered variables are independent of their non-descendants minus their parents by conditioning on their parents.

Causal Markov Condition (CMC) Pearl and Verma 1991

A probability distribution is compatible with a DAG G if, and only if, CMA is verified.

Consequences of CMC

Consequences:

- The joint density p verifies $p(x) = \prod_{i=1}^{d} p(x_i | x_{\mathrm{pa}_i})$.
- FCM $\Leftrightarrow$ causal graph.

Markov Blanket (MB)

For a given variable X_i, any minimal subset of the other variables such that any disjoint set of variables is independent of X_i conditioned on the subset is known as MB of X_i.

Moral graph

A moral graph of a DAG G is the undirected graph where each node is connected where the original node is connected with its MB in G.

Faithfulness

Causal Faithfulness

A graph G and a joint density $p(x)$ verify the Causal Faithfulness Assumption (CFA) if every Conditional Independence (CI) relation verified by p is entailed by G.

Causal Sufficiency

The Causal Sufficiency Assumption (CSA) states that the observed variables $X_1, \cdots, X_n$ are causally sufficient, *ie* each par of variables $\{X_i, X_j\} \subseteq \{X_1, \cdots, X_n\}$ do not has a common cause external to $\{X_1, \cdots, X_n\} \setminus \{X_i, X_j\}$.

Some basic three node structures

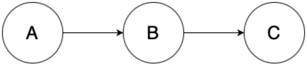

Figure 5: Example of a chain structure.

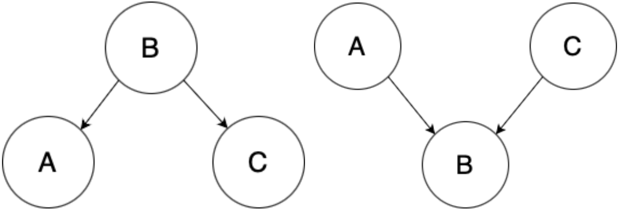

Figure 6: Example of a fork structure (left) and *v*-structure (right).

d-separation

A set of nodes W **blocks** a path t if

1. t contains at least one arrow-emiting node w $(i \rightarrow w \rightarrow j)$ or $(i \leftarrow w \rightarrow j)$ verifying $w \in W$.
2. t contains at least one collider node w $(i \rightarrow w \leftarrow j)$ verifying $w \notin W$ and not having any descendant on W.

The set W d-**separate** A and B in the graph G when W blocks all the paths between A and B.

- Under CFA: d-separation $\Leftrightarrow$ CI.

Markov Equivalence Class and CPDAG

Markov Equivalent DAG Pearl and Verma 1990

Two DAGs with same skeleton and same *v*-structures are said to be Markov equivalent.

Completed Partially Directed Acyclic Graph (CPDAG)

Graph with both directed and undirected edges representing a Markov Equivalence class.

Overview

Algorithms classification

Finding the causal graph only from observational data is a NP-hard problem Chickering, Heckerman, and Meek 2004.

Different basis algorithms:

1. Combinatorial: constrained and score-based: only able to find a CPDAG.
   - PC Spirtes, Glymour, and Scheines 2000.
   - GES Chickering 2002.
2. Continuous Optimization-based Approaches: find a DAG from distributional assymetries (usually based on additional assumptions).
   - NOTEARS Zheng et al. 2018: the first algorithm.
   - CGNN Goudet et al. 2018.
   - SAM.

SAM architecture

- GAN
  1. A generator for each variable
  2. An unique discriminator

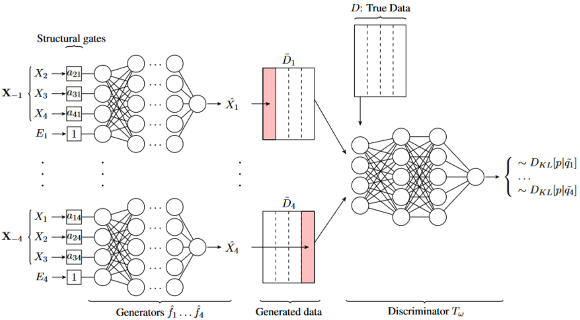

Figure 7: Architecture used in SAM (Figure from Kalainathan et al. 2022).

Optimization function

Aim: to minimize a loss combining both CI and distribution asymmetries.
Loss: based on the Markov Kernel of each variable.

$$\sum_{j=1}^{d}\left[\hat{I}^n(X_j, X_{\overline{Pa}(j;\hat{G})}|X_{Pa(j;\hat{G})})\right] + \lambda_S|\hat{G}| + \sum_{j=1}^{d}\left[\frac{1}{n}\sum_{l=1}^{n}\log\frac{p(x_j^{(l)}|x_{Pa(j;\hat{G})}^{(l)})}{q(x_j^{(l)}|x_{Pa(j;\hat{G})}^{(l)},\theta_j)} + \lambda_F||\theta_j||_F\right] + \lambda_D\sum_{k=1}^{d}\frac{\mathrm{tr}A^k}{k!}$$

- Structural gate matrix A.
- Structural loss: Identify the Markov Blanket of each variable.
- Parametric loss: Data fitting.
- Constraints: sparsity, causal mechanism power and acyclicity.

SAM key points

Theoretically:

- Structural gates as probabilities to make them differentiable Maddison, Mnih, and Teh 2017.
- Acyclicty constraint optimized through Augmented Lagrangian (AL) technique.
- It exists at least one positive value for the structural loss regularizer allowing the CPDAG identification by the minimization of the structural loss.

Experimentally

- Main benefit: versatility.
- Generally able to recover the true MB.
- Sensitive to the random initialization weights in the NN.

Overview

Why looking to the acyclicity constraint?

Motivation:

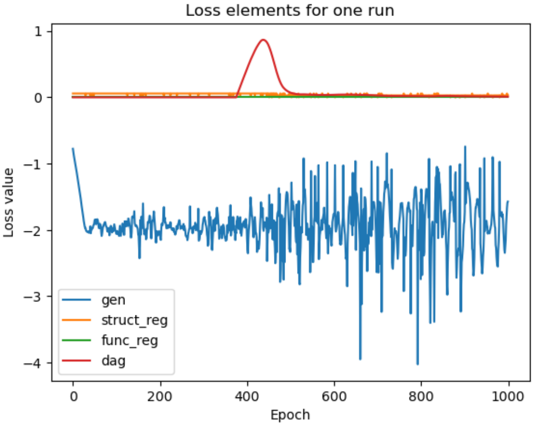

Figure 8: Evolution of the different loss elements, with a huge variability increment after the acyclicity constraint initialization.

Main question:

Is the stability affected by the AL optimization method?

In SAM:

- AL optimization penalization weight with additive increment.

My proposal:

- Consider a multiplicative increment on the AL.

Motivation

Main question:

- If $a_{i,j} = 0$, then it remains equal to zero?

My hypothesis:

- If $a_{i,j} = 0$ then it is probability of being selected is zero, then the NN is restricted to look for the space where it has another value.

Possible benefits:

- A way to encode prior knowledge of the non-existence of an edge.
- With the acyclicity assumption, a way to partially incorporate knowledge from the existence of an edge (through its reverse edge).

Initialization with a DAG

Main question:

- How does affect a DAG initialization to the random weights sensitivity?

Auxiliary questions:

- From the true graph?
- From the CPDAG?
- From adding an edge to the true graph?
- From removing an edge to the true graph?
- From reversing an edge to the true graph?

Possible benefits

- Analyze the interest of incorporating prior knowledge.
- Use SAM to test a solution through its stability.

Overview

Simulation setup

Datasets: 20 synthetic datasets based on Kalainathan, Goudet, and Dutta 2020.

- **Number of observations:** 1000.
- **Number of nodes:** 5.
- **Number maximum of parents**: 2
- **Mechanisms:** linear and neural network.
- **Noise :** $\mathcal{U}(0, 0.4)$, additive.

Evaluation measure: Standard deviation of the last epoch probabilities (without considering self-loops).

Methodology:

- 5 independent trials.

Numerical results

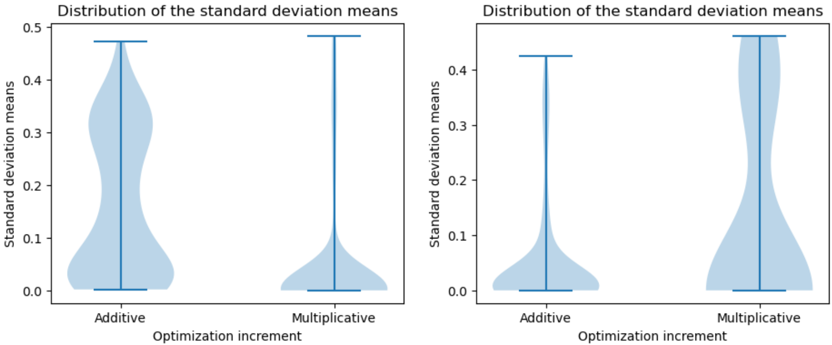

Figure 9: Impact of the increment for linear mechanism (left) and NN mechanism (right).

Simulation setup

Datasets: 10 synthetic datasets based on Kalainathan, Goudet, and Dutta 2020.

- **Number of observations:** 1000.
- **Number of nodes:** 2.
- **Mechanism:** linear.
- **Noise :** $\mathcal{U}(0, 0.2)$, additive.

Evaluation measure:

- Maximum value of any structural gate during the training.

Methodology:

- 5 independent trials.
- SAM parameters:
    - 750 epochs
    - $\Lambda_S = 0$
    - $\lambda_F = 2 \cdot 10^{-6}$.
    - $\lambda_D = 0.0$

Results

The structural gates value is always constant and equal to zero in all the graphs and independent trials.

Simulation setup

Datasets: 10 synthetic datasets based on Kalainathan, Goudet, and Dutta 2020.

- **Number of observations:** 1000.
- **Number of nodes:** 5.
- **Mechanisms:** linear.
- **Noise:** $\mathcal{U}(0, 0.4)$, additive.

Evaluation measure: Standard deviation of the last epoch probabilities (without considering self-loops).

Methodology:

- 5 independent trials.
- SAM parameters:
  - 1500 epochs
  - $\lambda_S = 0.02$
  - $\lambda_F = 2 \cdot 10^{-6}$.
  - $\lambda_D = 0.01$

Numerical results

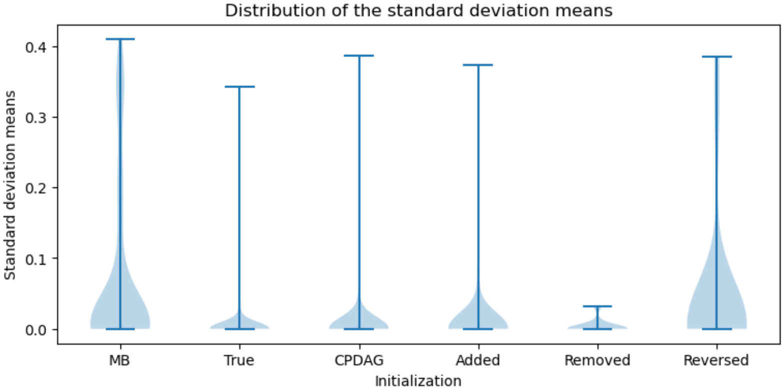

Figure 10: Violin plot for the standard deviation of the returned structural gates (without considering self-loops) for the different initializations.

Overview

Conclussion and discussion

AL penalized weight increment:

- The acyclicity constraint optimization method affects the stability.

Structural gates $a_{i,j} = 0$

- The structural gate seems constant if initialized to zero.

Initialize SAM with additional information

- One of the major sources of instability.
- The stability of SAM seems improved by all the initialization except when reversing an edge.

Further work

- Is SAM stability related to the causal mechanisms?
- Analyze the stability by pruning edges **during** the optimization procedure.
- Is the performance increased by considering an initialization obtained as a result from another causal discovery algorithm?

Thanks, any question?

References I

Chickering, David Maxwell (2002). "Optimal structure identification with greedy search". In: *Journal of machine learning research* 3.Nov, pp. 507–554.

Chickering, Max, David Heckerman, and Chris Meek (2004). "Large-sample learning of Bayesian networks is NP-hard". In: *Journal of Machine Learning Research* 5, pp. 1287–1330.

Goodfellow, Ian et al. (2014). "Generative adversarial nets". In: *Advances in neural information processing systems* 27.

Goudet, Olivier et al. (2018). "Learning functional causal models with generative neural networks". In: *Explainable and interpretable models in computer vision and machine learning*, pp. 39–80.

Kalainathan, Diviyan, Olivier Goudet, and Ritik Dutta (2020). "Causal discovery toolbox: uncovering causal relationships in python". In: *The Journal of Machine Learning Research* 21.1, pp. 1406–1410.

References II

Kalainathan, Diviyan et al. (2022). "Structural agnostic modeling: Adversarial learning of causal graphs". In: *The Journal of Machine Learning Research* 23.1, pp. 9831–9892.

Maddison, Chris J., Andriy Mnih, and Yee Whye Teh (2017). "The Concrete Distribution: A Continuous Relaxation of Discrete Random Variables". In: *International Conference on Learning Representations*.

Pearl, Judea (2003). "Causality: models, reasoning, and inference". In: *Econometric Theory* 19.4, pp. 675–685.

Pearl, Judea and T Verma (1991). *A theory of inferred causation. Principles of Knowledge Representation and Reasoning: Proceedings of the Second International Conference.*

Pearl, Judea and Thomas Verma (1990). *A Formal Theory of Inductive Causation*. Tech. rep. R-1555 (I). UCLA Cognitive Systems Laboratory.

References III

Spirtes, Peter, Clark N Glymour, and Richard Scheines (2000). *Causation, prediction, and search*. MIT press.

Zheng, Xun et al. (2018). "Dags with no tears: Continuous optimization for structure learning". In: *Advances in neural information processing systems* 31.

Greedy Equivalence Search (GES)

Pseudo-code:

- Select Bayesian Information Criterion (BIC) as scoring function
- Initialize an empty graph.
- Forward Equivalence Search:
  Find the best directed edge to add to the candidate CPDAG amongst all the missing edges. Repeat until score no longer improves.
- Backward Equivalence Search:
  Find the best directed edge to remove to the candidate CPDAG amongst all the present edges. Repeat until score no longer improves.
- Return directed graph.

Peter-Clark (PC) base algorithm

- Identify the skeleton:
  - Start with a complete graph
  - If $X \perp Y | Z$, remove edges $X - Y$ for some (initially empty) conditioning set Z and store Z as $Sepset(X, Y)$.
  - Repeat until possible by increasing the size of Z for each pair (X, Y).
- Identify v-structures and orient them.
  - For any undirected paths $X - Z - Y$, if $Z \notin Sepset(X, Y)$, then orient the undirected path as $X \rightarrow Z \leftarrow Y$.
- Orient qualifying edges that are incident on colliders.
  - For all $A \rightarrow B - C$, if A and C not adjacent then $B \rightarrow C$.
  - If it exists an undirected edge $A - B$ and a direct path from A to B, orient the edge as $A \rightarrow B$.

NOTEARS

Pseudo-code:

- Input: Initial guess (W_0, α_0), progress rate $c \in (0, 1)$, tolerance $\varepsilon > 0$, threeshold $\omega > 0$.
- Do:
  - Solve primal: $W_{t+1} \leftarrow \arg\min_W L^\rho(W, \alpha_t)$ with ρ such that $h(t_{W_{t+1}}) < ch(W_t)$.
  - Dual ascent: $\alpha_{t+1} \leftarrow \alpha_t + \rho h(W_{t+1}$.
  - If $h(W_{t+1}) < \varepsilon$, set $\tilde{W}_{\text{ECP}} = W_{t+1}$ and break.
- Threshold and return the matrix.