Stability Analysis of the Structural Agnostic Modeling Method

Research Internship Report

École Normale Supérieure Paris Saclay

M2 Mathématiques, Vision, Apprentissage

Ramón Daniel Regueiro Espiño
Supervisors: Alessandro Leite & Audrey Poinsot
Tutor: Remi Bardenet
Team: TAU (Inria - Saclay)

Paris, September 2023

ABSTRACT

A causal framework allows us to better understand the relationships between variables. Structural Agnostic Model (SAM) [Kal+22] aims to infer these relations from observational data. Nevertheless, its lack of stability restricts the confidence we can place in their outcomes. In this context, this report aims to study sources of SAM's stability and how to tackle them. The results highlighted that (a) giving an initialization strategy constitutes a major source of instability, and (b) initializing an edge as non-existent does partially help to handle it.

Contents

Chapter 1

Introduction

Although Machine Learning (ML) models can outperform humans on different tasks, they are usually unstable to varying types of noise, producing different predictions concerning small perturbations. This lack of stability makes it impossible to rely on the inferred knowledge as its performance might depend on an element different from the input data or the model parameters. Moreover, to replicate an unstable model might not be possible, which implies a lack of reproducibility of its results.

Besides the benefits of stability for ML models, most of the ML algorithms like linear regression or decision trees aim to extract correlation-based knowledge. However, these models can rely on spurious correlations, like it can be seen in Fig. 1.1, where the association between having yellow fingers and developing lung cancer without considering smoking. As a result of this, using these models as basis for interventions might lead to paint people fingers as a way to reduce the cases of lung cancer. Hence, inferring causal knowledge, i.e., understanding the causal-effect relations between variables, is useful in multiple real-world cases like social, behavioral, and health sciences. Causation helps us to understand the "how" and the "why" by answering questions like "what would happen if ...", complementing the knowledge inferred from correlation. For instance, it is proven that an association between low levels of vitamin D and some types of depression, but it is not clear if one is caused by the other [PBG17]. Hence, causality is needed to understand how we should intervene in the world, Public policy or medicine are examples of fields where a causal framework is crucial.

Randomized controlled trial (RCT) comprises the gold standard to infer cause-effect [IR15] relationships. A schema of an RCT can be seen in Fig. 1.2. In an RCT setting, individuals are randomly assigned either to a treated or a control group. This randomization allows to control of confounding variables and avoids the effect of external factors on the possible outcome.

However, RCTs are not always feasible due to temporal, economic, or ethical reasons. For instance, if we want to infer the effect of smoking on developing lung cancer we cannot force people to smoke to check it. Hence, interventional data, data where specific interventions are delivered and introduced, cannot always be used. In this case, we are only able to use semi-interventional data, where a degree of intervention is possible, or, in the worst case, observational data, where the data were collected without any deliberate manipulation. The

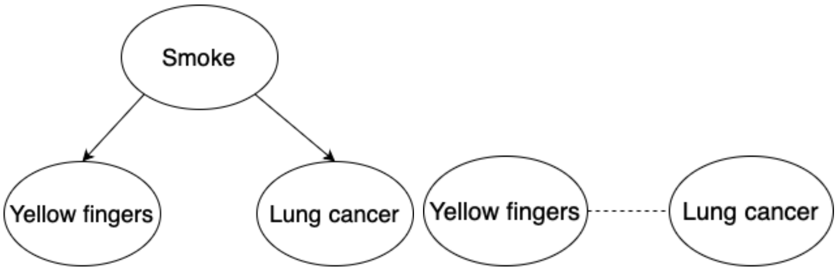

Figure 1.1: Example of causal relationship between smoking, having yellow fingers, and developing lung cancer (left) and spurious correlation relationship between having yellow fingers and developing lung cancer if smoke is not observed (right). Arrows represent causal-effect relationships ashed lines represent correlations

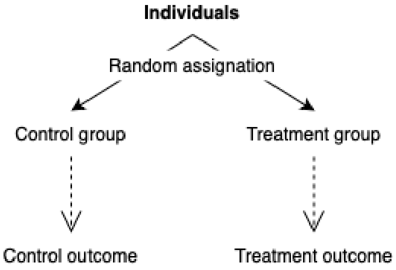

Figure 1.2: Schema of an RCT

ensemble of algorithms used to discover causal structures based only on observational data is known as *observational causal discovery*.

One example of an observational causal discovery algorithm is the Structural Agnostic Model (SAM) [Kal+22]. SAM aims to infer the underlying causal structure from data by leveraging both Conditional Independences (CI) and distributional asymmetries. For this, SAM uses an adversarial approach based on a Generative Adversarial Network (GAN) [Goo+14] architecture. SAM's main benefit in comparison with other causal discovery algorithms is its versatility under different scenarios. This property makes SAM extremely useful for real-world situations where we do not have prior knowledge like the causal structure or the form of the causal relations between variables.

Nevertheless, SAM is not a stable algorithm. This lack of stability limits is use and the reliability of the inferred causal relations. During the internship, we focus on analyzing if it can be stabilized under different conditions.

This analysis might allow us to understand the reasons for its instability. Furthermore, it can allow us to possibly handle them improving the performance and reliability of this causal discovery algorithm.

To better understand its instability we explored three different elements. Firstly, modifying the weight of a regularization term on the loss. Secondly, what happens if the initial probability of considering an edge is zero? Thirdly, how SAM stability is affected by giving an initial possible structure.

This internship report is organized as follows. Chapter 2 introduces the fundamental concepts of causal inference. Chapter 3 describes some related work on observational causal discovery. Chapter 4 presents some potential sources of SAM's instability. Chapter 5 describes the experimental results. Finally, Chapter 6 concludes this report and describes some future works.

Chapter 2

Foundation of Causality

The chapter aims to introduce some fundamental concepts of causality. Section 2.1 formally define causality under the interventionist viewpoint [Woo05]. Then, Sections 2.2 and 2.3 describe structural causal model and causal graph followed by the necessary assumptions in Section 2.5.

2.1 Causality

While there is no doubt about the definition of correlation, the notion of causation requires a deeper view which generates philosophical issues even nowadays [Pop22]. Among the literature, we highlight two different nature approaches to its description. Firstly, we can define causation based on the idea of what would have happened if things had been different [Lew73]. Secondly, we can define causation by considering probabilistic relations [Pea09].

The first approach considers causation as the dependency relation created by "A occurred and then B occurred" and "if A had not occurred, B would not have occurred" being verified [Kme20]. However, the only way to check this claim is to modify A by doing an action on it, i.e., intervening on A. In this case, if one can only change A without changing the other variables, we call this an *intervention*. An intervention enables one to answer questions "what would happen if it had happened A' instead of A?". This hypothetical scenario considering what would have happened if the circumstances were different by performing an intervention is known as *counterfactual* [Woo05; Pea09]. We can illustrate these notions through the relation between smoking and developing lung cancer. For this, we consider that a person smokes and he develops lung cancer. If we had perform an intervention on the smoking variable by enforcing them to not smoke, and them had not developed lung cancer we could conclude that smoking causes lung cancer.

However, we remark that both what is observed and the counterfactual cannot be observed at the same time, which is known as the "fundamental problem of causal inference" [Hol86]. In this case, we need to always consider the results based on causal hypotheses for a family of random variables with values that are never observed. Another limitation of the use of this approach is that usually we can show that B depends on A, supporting that A is a cause of B.

However, it is not easy to show that B does not depend on A in cases like where B can be both caused by A and A'.

As an alternative to the counterfactual approach, we can consider Pearl's approach to causation. Pearl's approach is based on analyzing the changes on the distribution of a variable Y in function of an intervention on the variable X through the $do(x)$ operator. This operator consists on artificially replacing X for a fixed and constant value x without changing the other elements of the model. Then, we can compare the different distributions of Y with relation to the different values which X was replaced by.

2.2 Structural Causal Model

The main idea of Pearl's approach can be expressed through a mathematical representation known as Structural Causal Model (SCM) [Pea09].

Definition 1. *Considering a set of random variables $X_1, \ldots, X_d$, a SCM is a pair (S, P_N) where S is a collection of d structural assignments*

$$X_j = f_j(Pa_j, E_j), \quad j = 1, \ldots, d,$$

where $Pa_j \subseteq \{X_1, \ldots, X_d\} \setminus \{X_j\}$ are called parents of X_j; and a joint distribution $P_E = P_{E_1}, \ldots, P_{E_d}$ over the noise variables E_j which we require to be independent.

Historically, one of the initial SCMs is the Structural Equation Model (SEM). The SEMs has been used during the last decades in different fields like sociology [BH77], marketing [Ric+16], or ecology [Fan+16], to represent causal linear relations between variables.

Definition 2. *Considering a set of random variables $X_1, \ldots, X_d$, a SEM is a set of equations*

$$X_i = \sum_{j \neq i} a_{ij} X_j + E_i \quad i = 1, \ldots, d, \tag{2.1}$$

where, E_i represents the random noise modeling the effects of non-observed factors and the coefficients $a_{i,j}$ represent the expected effect of X_j on X_i.

From a considered SEM, we highlight that two variables X_j and X_i hold a cause-effect linear relationship if, and only if, $a_{ij} \neq 0$ in the corresponding Eq. (2.1). SEMs are based on linear parameters and cannot express all the different nonlinear real-world causal relations [BP13]. The generalization of the SEM to express the non-linear case is known as the Functional Causal Model (FCM).

Definition 3. *Considering a set of random variables $X_1, \ldots, X_n$, a FCM is a set of equations*

$$X_i = f_i(Pa_i, E_i) \quad i = 1, \ldots, n \tag{2.2}$$

where, each f_i is a function called a causal mechanism.

An example of FCM for a set of variables $X_1, \ldots, X_5$ can be seen in Figure 2.1. In this FCM, the variables X_1 and X_4 do not have any parents so they are directly determined by their

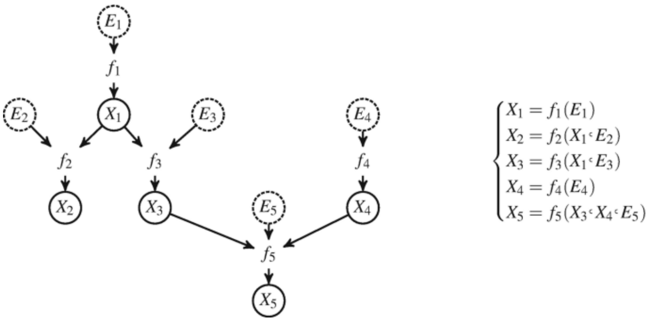

Figure 2.1: Example of a functional causal model. Dashed circles represent exogenous variables, non-dashed ones endogenous, and arrows direct cause-effects [Gou+18]

correspondent exogenous variables. Moreover, the other variables X_2, X_3 and X_5 are causally determined by pa_2 and E_2, pa_3 and E_3 and pa_5 and E_5 respectively.

2.3 Causal Graph

Causal graphs are a visual way to encode causal structures. In a causal graph, each node represents a variable and each directed edge indicates a cause-effect relation between the variables represented by these nodes, from the cause to the effect. Usually, for simplicity, the exogenous variables are omitted from the graph. An example of a causal graph can be seen in Fig. 1.1. On the left, we can see a causal graph relating smoking, designated by the node *smoking*, as a common cause of having yellowish fingers, represented by the node *yellow fingers*, and developing lung cancer, denoted by the node *lung cancer*.

A remark related to causal graphs is that they do not correspond to the same concept as Bayesian networks. For example, Bayesian networks can represent relations that are not causal. Moreover, causal graphs can present cycles. However, dealing with cycles is challenging and requires many more assumptions to perform causal inference. That is why, we focus our work on causal graphs which do not contain any direct cycles. This type of graph is known as Directed Acyclic Graphs (DAG).

2.4 Markov Assumption, d-separation and Faithfulness

Once we have introduced both ways of representing causal relations, FCMs, and causal graphs, we can ask ourselves if a connection can be established between both of them. In this section, we explore the connections between distributions, FCMs, and DAGs.

As a first remark, we highlight that we do not have presented any theoretical result that ensures the relation between the causal graph and the probability distribution associated to the nodes generated by the graph. To solve this issue, we assume that the distribution matches the graph, which is known as the Causal Markov Assumption (CMA).

Definition 4. *Given a probabilistic distribution $P(X_1, \ldots, X_d)$ we say that the associated DAG G verifies the Causal Markov Assumption if all the considered variables are independent* *of their non-descendants minus their parents by conditioning on their parents*

$$\forall j[1,d] \in,\ X_j \perp\!\!\!\perp ND_j \setminus Pa_j | Pa_j,$$

where $Pa_j, ND_j \subseteq \{X_1, \ldots, X_d\} \setminus \{X_j\}$ denotes the parents and the non-descendants nodes of j in the graph.

Given the joint density p and the associated DAG G, we have that under CMA p factorizes as follows:

$$p(x) = \prod_{i=1}^{d} p(x_i|\mathrm{Pa}_i). \tag{2.3}$$

Under the CMA, we can deduce CI based on the factorization described in Eq. (2.3). Furthermore, we can infer dependencies based on the Causal Faithfulness Assumption (CFA).

Definition 5. *The CFA states that the joint density p is assumed to be faithful to the graph G, that is, every conditional independence relation that holds true according to p is entailed by G.*

We highlight that by considering only the CFA we might not be able to infer all the dependencies. For instance, if an unobserved common cause exist between two variables. In Figure 1.1 we found a dependency that is not explained in the graph is the variable Smoke is not observed. To ensure that all dependencies are explained by a path on the causal graph, we need to assume that unobserved common causes do not exist. This assumption is known as the Causal Sufficiency Assumption (CSA).

Definition 6. *The CSA states that the observed variables $X_1, \ldots, X_n$ are causally sufficient, i.e., each par of variables $\{X_i, X_j\} \subseteq \{X_1, \ldots, X_n\}$ do not has a common cause external to $\{X_1, \ldots, X_n\} \setminus \{X_i, X_j\}$.*

From the FCM point of view, the CSA corresponds to assuming that the exogenous variables are independent of each other.

Hence, if the three assumptions, CMA, CFA and CSA, are verified, we can deduce dependencies and independencies based on the observed causal graph. For this, we use a set of tools known as d-separation techniques which are based on the graphical concept of path.

Definition 7. *A path t in a graph is a sequence of at least two distinct nodes $i_1, \ldots, i_m$ such that there exists a direct edge from i_k to i_{k+1} or from i_{k+1} to i_k for all $k = 1, \ldots, m-1$.*

A path t in a graph G is said to be d-separated or blocked by a set of nodes W if

1. *t contains at least one arrow-emitting 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.*

If all the paths between two sets of nodes A and B are blocked by a third set of nodes W, we say that the set W d-separates A and B in the graph G.

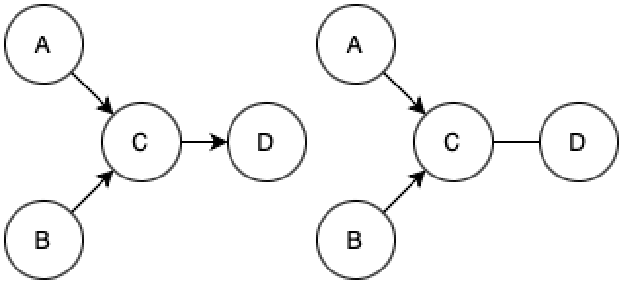

Figure 2.2: Example of a DAG (left) and the associated graph combining its skeleton and v-structures (right)

Hence, under the CMA, CFA and CSA, we have that the set of nodes A and B are CI by conditioning on the set of nodes W if, and only if, A and B are d-separated by W.

As a consequence of applying d-separation techniques we can deduce the minimal set of nodes needed to encode all the information related to the distribution of a variable X_j. For a fixed node this minimal set includes all the parents of the node, the children of the node and the parents of its children. Moreover, it is known as the Markov Blanket (MB) of the node.

Definition 8. *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.*

Based on a given DAG G, the undirected graph where each node is connected to its MB is known as moral graph.

2.5 Markov Equivalence Class and Completed Partially Directed Acyclic Graphs

Once we introduce the d-separation set of techniques and how it helps to identify some causal relations, we can ask if all the different DAGs can be identified through this set of techniques. On the one hand, we can obtain the skeleton and the v-structures. Also, we can identify the orientation of some additional edges from s skeleton and v-structures based on knowing that they are not part of a v-structure, like the edge $C - D$ in Figure 2.2. On the other hand, not all causal relations can be identified through CIs. For instance, we cannot make any distinction between the chain and the fork based only on CI tests.

Indeed, all the causal graphs that verify the same CIs belong to the same equivalence class, known as Markov Equivalence Class [PV91].

Property 1. *Two DAGs with the same skeleton and v-structures belong to the same Markov Equivalence Class.*

The representative of each Markov Equivalence Class is known as a Completed Partially Directed Acyclic Graph (CPDAG).

Definition 9. *A graph with both directed and undirected edges representing a Markov Equivalence class is known as CPDAG.*

Chapter 3

Causal Discovery

Finding the causal graph just based on observational data is an NP-hard problem [Chi96; CHM04]. To tackle this problem, algorithms based on different approaches can be considered to infer the causal structure. We can classify them between combinatorics and continuous optimization-based approaches.

3.1 Constrained and Score-based Methods

Until the proposal of the first differentiable acyclicity constraint [Zhe+18], the main basis of causal discovery algorithms was a combinatorics approach enforcing the acyclicity of the graph. Mainly, the algorithms based on this type of search are based on constraints, on optimizing a score or a combination of both of them, by optimizing a score on a space restricted by conditional independence (CI) tests.

Constraint-based approaches rely on the relation of CI tests and d-separation techniques to infer information about the causal graph. Nevertheless, as we mentioned before, different DAGs can hold the same CIs. Remarkably, we have that these algorithms cannot answer the bivariate case, where only two nodes are present and the aim is to orient the present edge. An example of a CI-based algorithm is the Peter-Clark (PC) [SGS00] algorithm.

The PC algorithm aims to return a Partially Directed Acyclic Graph (PDAG) by assuming acyclicity, CMA, CFA, and CSA. This algorithm has three different phases. Firstly, it recovers the skeleton of the causal graph from the CI tests. Secondly, it uses the collider nodes relation with CI to identify them and obtain the corresponding CPDAG. Thirdly, considering each of the remaining undirected edges, orient the considered undirected edge if, in the other sense, it creates a new v-structure or a cycle.

In contrast to constrained-based algorithms, score-based algorithms aim to test how well each graph structure fits the data. For this, if A denotes the adjacency matrix of the graph, they aim to solve the optimization problem described in Eq. (3.1). Usually, the score S penalizes the wrong conditional independences which can yield to bad model fitting combining it with an element that penalizes the complexity of the model. As an example of a score, we highlight the Bayesian Information Criterion (BIC), which is the chosen score in algorithms like the Greedy Equivalence Search (GES) [Chi02].

$$\begin{aligned} &\min_{A \in \mathbb{R}^{d \times d}} S(A) \\ &\text{subject to } G(A) \in \text{DAGs} \end{aligned} \tag{3.1}$$

GES is a two-stage algorithm based on the operation of adding an edge, corresponding to the Forward Equivalence Search phase, and the operation of removing an edge, corresponding to the Backward Equivalence Search phase. From the empty graph, it starts by adding the possible direct edges which increase the most the BIC until any addition does not increase the score. Then, it removes the edge which resultant model increases the most the BIC until a local maximum is achieved. Then, it returns the obtained DAG. From a theoretical point of view, under the CMA, CFA, and CSA it can recover a graph with the same CPDAG as the true causal graph.

3.2 Continuous Optimization-based Approaches

Continuous optimization-based approaches aim to infer the causal structure by optimizing a differentiable score, usually by the use of Deep Learning (DL). For this, instead of considering the problem described in Eq. (3.1) they consider a constraint considering a differentiable function h as it is described in the Eq. (3.2).

$$\begin{aligned} &\min_{A \in \mathbb{R}^{d \times d}} S(A) \\ &\text{subject to } h(A) = 0 \end{aligned} \tag{3.2}$$

3.2.1 NO TEARS

Non-combinatorics Optimization via Trace Exponential Augmented Lagrangian Structure learning (NO TEARS) [Zhe+18] is considered as the first algorithm that overcomes the use of a combinatorics approach to characterize a graph [VCB22]. For this, it uses an acyclicity constraint described in the Eq. (3.3) to infer a causal graph based on linear mechanisms.

$$h(A) = \mathrm{tr}(e^{A \odot A}) - d = 0, \tag{3.3}$$

where $\odot$ indicates the Hadamard product. While it allows us to consider a differentiable approach by considering a weighted adjacency matrix, we highlight that solving this constraint requires an $O(d^3)$ algorithm [AH10].

The proposed NO TEARS algorithm uses a dual ascent optimization method to solve a sequence of unconstrained problems instead of the constrained problem obtained from an Augmented Lagrangian (AL) approach. After several iterations, it does a threshold on the weighted adjacency matrix fixing to zero the values smaller than a considered hyper-parameter and returns the graph obtained from the weighted adjacency matrix.

3.2.2 Causal Generative Neural Network

A different approach to the NO TEARS algorithm which seeks to estimate the full DAG and allows other mechanisms than linear is the Causal Generative Neural Network (CGNN) [Gou+18] framework.

This generative model aims to minimize the Maximum Mean Discrepancy (MMD) from the real distribution to the estimated distribution $\widehat{P}$ from $(\widehat{G}, \widehat{f}, \varepsilon)$.

From the given skeleton of G, it starts by considering each undirected edge and orienting it. Afterward, it adjusts each graph to be a DAG by finding each possible cycle and reversing an edge from each cycle to remove them. Finally, for a fixed number of iterations, it reverses the edge that minimizes the MMD without creating a new cycle.

We highlight that this algorithm can deal with hidden confounders. For this, based on correlations between exogenous variables, if it finds a hidden confounder between two variables it introduces an additional exogenous variable as a common cause of both variables. However, we highlight that CGNN has a high computational cost, quadratic complexity, and a scalability issue due to the use of a greedy search.

3.2.3 Structural Agnostic Modeling

Similarly, SAM's algorithm aims to extract a DAG $\widehat{G}$ by estimating the causal mechanisms $\widehat{f} = (\widehat{f}_1, \cdots, \widehat{f}_d)$ from a set of continuous variables. To reduce the limitation of the computational cost associated with CGNN it uses an adversarial model, a GAN, to estimate the loss. Moreover, instead of minimizing the MMD, it optimizes a function based on the sum of log-likelihoods of the conditional distribution of each variable.

Each directed graph is estimated through its characterization by its adjacency matrix. In this case, each column $a_j = (a_{1,j}, \cdots, a_{d,j})$ is a binary vector, known as **structural gate**, where $a_{k,j} = 1$ denotes that X_k is a cause of X_j and $a_{k,j} = 0$ otherwise. To avoid self-loops, each coefficient $a_{j,j}$ is fixed to zero. Moreover, each coefficient $a_{i,j}$ is made differentiable through a tool known as the Bernoulli reparametrization trick [MMT16]. This trick consists of considering each of these discrete binary variables as the parameter of a Bernoulli distribution which indicates if the considered edge belongs to the candidate graph $\widehat{G}$.

As we mentioned before, the chosen model architecture is a GAN with a different generator for each variable j as can be seen in Figure 3.1. The aim of each of these generators is to replicate the conditional distribution q_j of the variable. Moreover, the impact of the exogenous variable E_j is always considered for generating the conditional distribution of each variable X_j. To be able to infer different causal mechanisms, each conditional distribution is modeled as an NN.

Based on a unique observation, the set of generators $\widehat{f} = (\widehat{f}_1, \cdots, \widehat{f}_d)$ generates a new synthetic observation. After, a unique discriminator, modeled as an NN aims to distinguish between the synthetic and the true observations to estimate the loss described by:

$$\sum_{j=1}^{d}\left[\hat{I}^n(X_j, X_{\overline{Pa}(j;\widehat{G})}|X_{Pa(j;\widehat{G})})\right] + \sum_{j=1}^{d}\left[\frac{1}{n}\sum_{l=1}^{n}\log\frac{p(x_j^{(l)}|x^{(l)}_{Pa(j;\widehat{G})})}{q(x_j^{(l)}|x^{(l)}_{Pa(j;\widehat{G})}, \theta_j)}\right], \qquad (3.4)$$

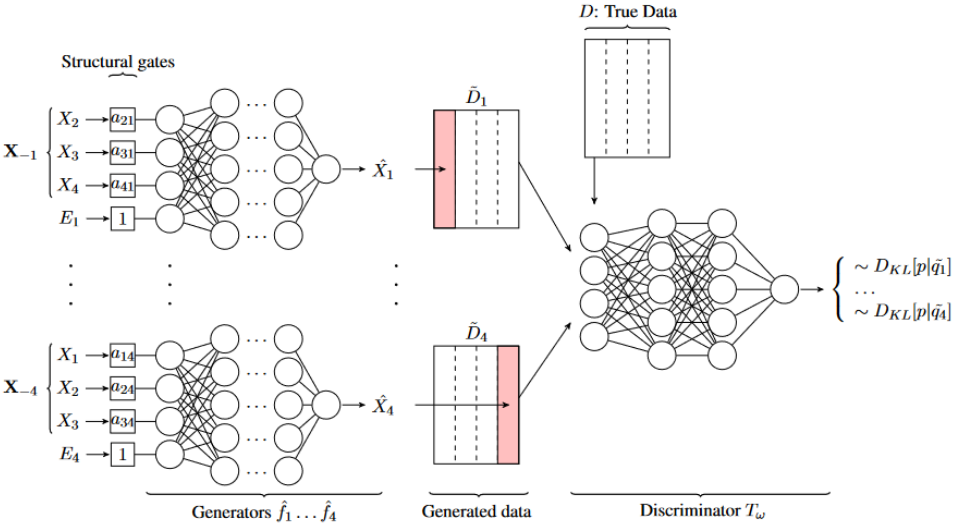

Figure 3.1: SAM architecture for a four variable dataset [Kal+22]

where $\hat{I}^n(X_j, X_{\overline{Pa(j;\widehat{G})}}|X_{Pa(j;\widehat{G})})$ denotes the empirical conditional mutual information term of each variable conditioned on $\overline{\mathrm{Pa}(j;\widehat{G})}$, which are the set of variables distinct to j and its parents on the graph $\widehat{G}$. We highlight that the Mutual Information (MI) quantifies the amount of information, through its entropy [Sha48], obtained from one random variable by observing the other one. In addition to this, we remark that two random variables X_i and X_j are CI from a third variable X_z if, and only if, $I(X_i, X_j|X_z) = 0$.

However, we highlight that from the MI we are, at best, only able to recover the minimal set of nodes giving information on the considered variable X_j, *ie* its MB. To not consider the edges between X_j and the other parents of its children a regularization term, $\lambda_S|\widehat{G}|$, proportional to the number of edges of the candidate graph is added. The key idea is that, by adding this term controlling the sparsity of the graph, SAM should be able to detect when an edge is added only because both nodes have a common child and then delete it. The sum of both terms, the empirical conditional mutual information term, and the regularization term, is known as structural loss and it is shown theoretically that, by a right choice of λ_S, the true CPDAG can be identified by minimizing this loss.

Theorem 3.2.1 (CPDAG identification by structural loss minimization)**.** *Under CMA, CFA and CSA assumptions, two results of convergence in probability, hold:*

1. *For every DAG* $\widehat{G}$ *in the equivalence class of* G, $\lim_{n\to\infty} \mathcal{L}_S^n(\widehat{G}, D) = \mathcal{L}_S^n(G, D)$.

2. *For every DAG* $\widehat{G}$ *not in the equivalence class of* G*, there exists* $\lambda_S > 0$ *such that:* $\lim_{n\to\infty} \mathbb{P}(\mathcal{L}_S^n(\widehat{G}, D) > \mathcal{L}_S^n(G, D)) = 1$.

In addition to the empirical conditional mutual information term, the other term of the Eq. (3.4) aims to infer the conditional asymmetries by measuring the ability of the generator, taking into account the considered structural gates to fit the conditional distribution. Moreover, to restrict the power of the causal mechanisms, which can lead to an overfitting and to be able to consider a $\widehat{G} \neq G$, it includes a regularization term $\lambda_F||\theta_j||_F$ based on the Frobenius norm. The sum of both terms controls the estimated mechanisms and it is known as parametric loss.

Finally, a third term is based on the acyclicity constraint introduced in NO TEARS and described in the Eq. (3.3).

$$\sum_{k=1}^{d} \frac{\mathrm{tr} A^k}{k!} = 0. \tag{3.5}$$

The main point behind the constraint of the Eq. (3.5) is that the k-th power of this matrix, the element (i, j) is strictly positive if, and only if, a k-length path from node i to j exists. So, if there are no cycles of length k, then $\mathrm{tr}(A^k) = 0$.

To avoid the combinatoric search by using this constraint, the considered problem is transformed into a sequence of unconstrained problems using an AL optimization method. In this case, the penalization term is expressed as $\lambda_D \sum_{k=1}^{d} \frac{\mathrm{tr} A^k}{k!}$, with $\lambda_D = 0$ until a certain epoch, which allows to find the MB of each node. Then, it has an additive increment of a fixed value per epoch to penalize the existence of cycles in the graph.

To conclude, we remark that by adding to the Eq. (3.4) these three regularization terms described before we obtain the final loss which aims to be optimized. This loss is

$$\sum_{j=1}^{d} \left[\widehat{I}^n (X_j, X_{\overline{Pa(j;\widehat{})}} | X_{Pa(j;\widehat{G})}) \right] + \lambda_S |\widehat{G}| + \sum_{j=1}^{d} \left[\frac{1}{n} \sum_{l=1}^{n} \log \frac{p(x_j^{(l)} | x_{Pa(j;\widehat{G})}^{(l)})}{q(x_j^{(l)} | x_{Pa(j;\hat{G})}^{(l)}, \theta_j)} + \lambda_F ||\theta_j||_F \right] + \lambda_D \sum_{k=1}^{d} \frac{\mathrm{tr} A^k}{k!}. \tag{3.6}$$

Chapter 4

Sources of Instability of the Structural Agnostic Model Method

While SAM tackles the computational cost of CGNN by adversarial learning, its lack of stability is a main limitation. This notion can be described in different ways, for instance, by the uniform stability [BE02].

Definition 10. *If $\mathcal{Z}$ denote the product of the input and output spaces of algorithm A and S a training set, we say that A has uniform stability β concerning the loss function l if*

$$\forall S \in \mathcal{Z}^m, \forall i \in \{1, \cdots, m\}, ||l(A_S, \cdot) - l(A_{S^{\setminus i}}, \cdot)||_\infty \leq \beta. \tag{4.1}$$

Moreover, we focus on the case where the algorithm is randomized, which is SAM's case. For this, we can consider the generalization of the equation (4.1) provided by [HRS16] where it was applied to the stochastic gradient descent (SGD) algorithm.

Definition 11. *A randomized algorithm A is β-uniformly stable if $\forall S, S' \in \mathcal{Z}^m$ such that S and S' differ in at most one example, we have*

$$\sup_z \mathbb{E}_A[l(A(S); z) - l(A(S'); z)] \leq \beta. \tag{4.2}$$

We highlight that, while in the equation (4.1) the variability on the loss is only considered based on the training dataset the equation (4.2) considers also the possibility of having different results based on the same training data. In our concrete case, this can be seen when SAM is initialized with the same training dataset but a different pair $(\widehat{G}, \widehat{f})$ is obtained due to the random initialization of the NN weights.

In this chapter, we study some possible elements that might impact its lack of stability, and that could guide future research lines to improve SAM's stability.

4.1 Impact of the acyclicity optimization method

From a theoretical point of view, if the right λ_S is chosen the true CPDAG can be found by minimizing the structural loss. From the experimental results of [Kal+22], SAM usually finds the moral graph if the acyclicity constraint is not added. Concerning this, from Figure 4.1 we can imagine that the acyclicity constraint might increase also the variability of the loss. This huge variability might come from the discriminant error of the GAN and it might result in not being able to correctly minimize its value. Hence, a first axe of work might be how the use of $\lambda_D \sum_{i=1}^{d} \frac{\text{tr}(A^k)}{k!}$ can impact the stability of the results.

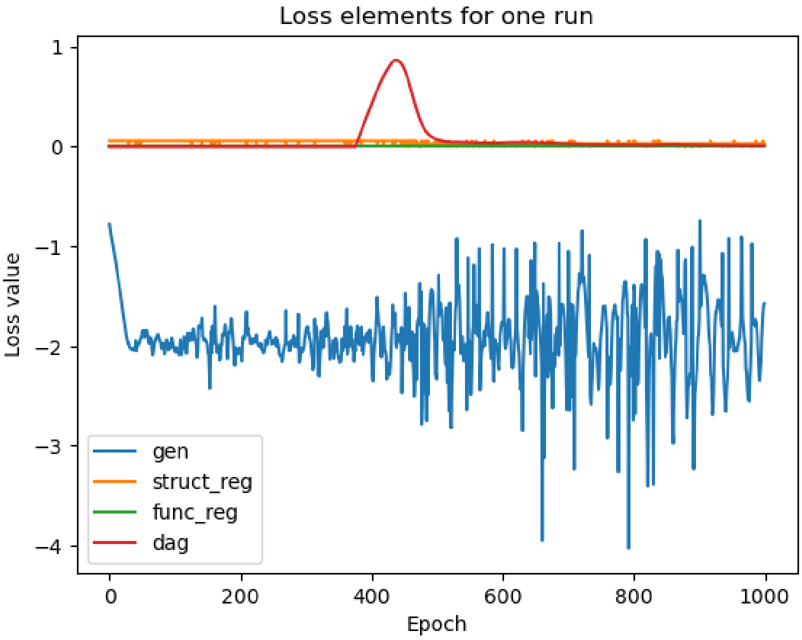

Figure 4.1: Evolution of the different loss elements for one training, with a huge variability increment after the acyclicity constraint initialization. *gen* refers to the discriminator error described in equation (3.4), *struct_ reg* denotes the sparsity regularization term, *func_ reg* indicates the mechanisms regularization term and *dag* refers to the value of the acyclicity regularization term.

A possible explanation of this is that, by considering an AL approach, we progressively reduce the interest in exploring some possible edge reversals, as they can create a cycle, except in the case that they might produce a huge impact on the estimated loss. However, as we mentioned before, this impact can be produced by the discriminator error and it might not be related to a true variation on the true value of the equation (3.6).

A possible solution to explore might be to consider different acyclicity constraints like the one introduced in Gradient-based Optimization of dag-penalized Likelihood for learning linEar dag Models (GOLEM) [NGZ20], which requires only a soft constraint. However, I decided to focus on the optimization method of SAM.

For this, I proposed to use a multiplicative increment instead of an additive increment on the AL whose experimental results are available in Section 5.1. The core is that, during the first epochs where it is considered, it considers a smaller weight than the additive increment. Hence, it might allow a better initial exploration of the edges' orientation. In addition, after this initial exploration, the parameter should strongly increase, enforcing the DAG $\widehat{G}$ to have always the same edges and better exploring the mechanisms associated with $\widehat{G}$. Hence, this approach might avoid possible erroneous estimations of the mechanisms because the weights were optimized for different structural gates. We remark that, while on the first layer, this might not be crucial, it might be more important on the hidden layers where the interaction between different observed variables can play a more important role.

As a main limitation of this approach, I consider that the increment multiplicative parameter should be carefully chosen in function of the number of epochs. Specially, from a practical point of view too high values should be avoided. Otherwise, the acyclicity constraint term of the loss might not be handled by the available computational resources and then they might be computed as infinite.

4.2 Fixing the coefficients $a_{i,j}$ to zero

While in [Kal+22] they state that each structural gate, except the self-loops, is initialized to zero with probability 0.5 I considered what happens if a $a_{i,j}$ is initialized to zero. In this case, based on the Bernoulli reparametrization trick, if a structural gate is fixed to zero this implies that this edge should not be never considered as part of this graph. As a way to reinforce this hypothesis, we have that this is what is considered to avoid the effect of self-loops. In this case, my intuition was that initially considering $a_{i,j} = 0$ implies that $a_{i,j}$ will always be constant and equal to zero. The associated experimental results can be seen in Section 5.2.

We highlight that, if this hypothesis is verified, it can be seen as a way to constrain SAM to respect the non-existence of a causal link. In addition, by considering the acyclicity assumption, if we know the existence of an edge $X_i \to X_j$, by fixing $a_{j,i} = 0$ we can partially encode the information related to knowing the existence of a causal link. Moreover, the only need of SAM is to capture the existence of the edge, which theoretically should be done for a good value λ_S by the CPDAG identification by structural loss minimization theorem.

4.3 Initialization of the structural gates

A possible way to reduce the impact of the random weights sensitivity is to give an initial graph to SAM. In this case, the random initialization weights might converge to the weights that might optimize the loss for each conditional distribution and, as the edges do not need to be oriented, they cannot affect the stability. Moreover, the weights related to the $a_{i,j}$ initialized as zero should be zero due to the penalization of the expressive power of the mechanism. Hence, we ask ourselves if this might not be false by considering different initializations of graphs.

The verification of this hypothesis might allow us to have a clue about the impact of incorporating prior knowledge. Moreover, it will allow us to consider the interest in developing a two-step algorithm where in the first phase an initial guess of the causal structure is done. Then, in a second phase, one run of SAM is done to recover the final pair $(\widehat{G}, \widehat{f})$. For this, we ask how the stability of SAM evolves by considering different knowledge of the causal structure: the MB, the CPDAG, and the true DAG. Additionally, we ask if SAM can infer that false knowledge is stable when it is provided. The experimental results related to these questions can be found in Section 5.3.

Chapter 5

Experimental results

In this chapter, we collect the results of the experiments done to test the hypothesis described in Chapter 4. Firstly, we introduce the experimental results related to the impact of the penalized weight increment on the acyclicity constraint. Secondly, we study if the initialization of the structural gates to zero forces them to be constant. Thirdly, we study the impact of considering different graphs as initializations on the stability of SAM.

To perform all the experiments, we consider synthetic datasets generated based on SCMs using the class *AcyclicGraphGenerator* from the Causal Discovery Toolbox [KGD20]. Each DAG was generated given a fixed number of nodes and a fixed maximum number of parents for the nodes. In case a DAG did not contain any edge it was automatically generated again. The nodes without parents were generated using Gaussian Mixture Models (GMMs) with four components and spherical covariance. In addition, noise variables were added to the SCMs. Each generated dataset was i.i.d. and each variable was obtained from the computations of the generated mechanisms following a topological order of the causal graph.

SAM's architecture parameters are fixed for all the experiments. Each generator is considered as two hidden layers of twenty neurons each with activation function tanh. The discriminator is an NN with two hidden layers of two hundred neurons with LeakyReLU as an activation function and batch normalization. The training is done using Adam [KB15].

5.1 Impact of the acyclicity optimization method

To evaluate the impact of the acyclicity optimization method on stability, we consider 20 graphs, each consisting of 5 nodes with a maximum of 2 parents per node. The considered scenarios were 10 graphs with linear mechanisms and 10 graphs with NN mechanisms. For each causal graph, we consider an additive noise following the distribution $\mathcal{U}(0, 0.4)$ and we generate 1000 observations. We ensure that at least one structure $A \to B \leftarrow C$ and one structure $A \leftarrow B \to C$ were generated in each scenario. The number of nodes of each graph was chosen to better assess the understanding of each possible structure on SAM's stability.

SAM hyper-parameters were chosen based on the experimental settings of [Kal+22] and the ones rendering better performance were kept. SAM was trained during 1500 epochs, with

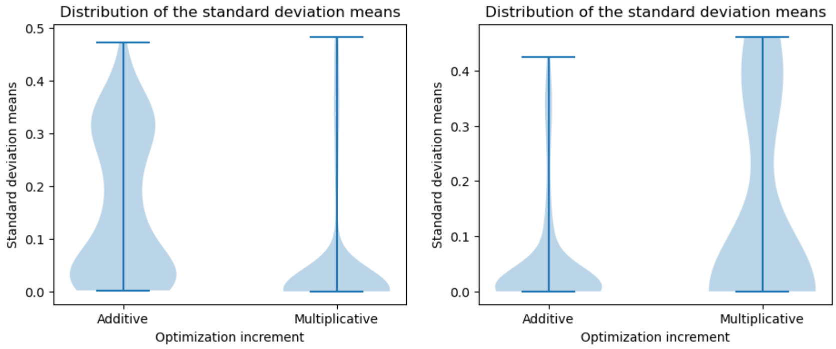

Figure 5.1: Violin plots showing the impact of the different ways of increment for linear mechanism (left) and NN mechanism (right) on SAM's variability.

$\lambda_S = 0.02$, and $\lambda_F = 2 \cdot 10^{-6}$. On the one hand, if the optimization increment is additive, we consider $\lambda_D = 0.01$. On the other hand, if it is multiplicative, we consider the increment as $\lambda_{D,\text{mult}} = 1.01$ to prevent the possible computational explosion of this weight value.

For each graph, we conduct five independent trials and use as a metric the mean of the standard deviation of the returned probabilities of each coefficient $a_{i,j}$ where $i \neq j$. These coefficients can be interpreted as the probability of each edge for being selected on the returned graph. It is important to remark that this metric is beneficial compared to one based only on the returned graph as it also benefits from the exact numerical value of each coefficient instead of operating to select the edges from the weighted matrix of structural gates.

An additional metric that was initially considered is the Structural Hamming Distance (SHD). This distance is defined as the minimum number of edge additions, removals, and reversals required to transform one graph into another. This metric has already been used to asses the stability of causal discovery methods [CM+14]. Its main benefit can be considering the differences between misorienting an edge in relation with adding a non-existent or removing an existent edge from the graph. As its main limitation to assess SAM's stability we consider the loss of information related to the transformation of structural gates into a graph and the bias introduced by this process. However, we finally did not consider it given the small number of nodes of the considered graphs.

As it can be seen in Fig. 5.1, it seems that the chosen way to increment the AL optimization method impacts the variability. However, both ways show different performance results based on the causal mechanisms. While an additive increment results in worst performance for a linear mechanism, it shows better performance for a NN mechanism. Moreover, it seems that the stability for fixed hyper-parameters is different if the mechanism nature is different.

5.2 Fixing the coefficients $a_{i,j}$ to zero

To experimentally check if fixing the coefficients of the structural gates to zero implies that they remain to zero we consider 10 graphs from the bivariate case. This concrete case is a causal structure of only two variables where one is the cause and the other the effect. As

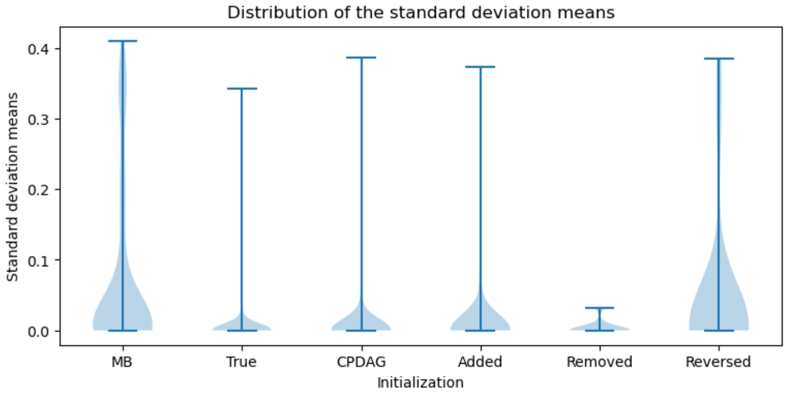

Figure 5.2: Violin plot for the means of the standard deviation of the returned $a_{i,j}$ (without considering self-loops) for the different initializations.

in the previous case, we consider 1000 observations for each graph and we consider a linear mechanism since it requires a smaller expressive power to be learned. Moreover, to reduce the impact of a noise we consider an additive noise following a distribution $\mathcal{U}(0, 0.2)$.

For each graph, five independent trials are considered. SAM's chosen parameters are 750 epochs and $\lambda_F = 2 \cdot 10^{-6}$. The weights of the sparsity and acyclicity regularization terms, Λ_S and λ_D, are fixed to zero as a way to avoid any possible augmentation on the loss value produced by an increment of the coefficients of the structural gates.

As a way to evaluate the experiments, we consider measuring the maximum value of any coefficient during all the training epochs.

The empirical results of all the different trials are that the maximum value of any $a_{i,j}$ for any epoch is constant to zero. Hence, it seems that fixing these coefficients to zero might imply that the edges are never considered and the initial hypothesis is not rejected.

5.3 Initialization of the structural gates

As a way to check the impact of giving an initial structure on SAM and how it affects its stability, we consider the 10 graphs with linear mechanisms described in Section 5.1. Moreover, as we want to analyze the same property, SAM's stability, we consider the same evaluation measure due to the same reasons. SAM's parameters are 1500 epochs, $\lambda_S = 0.02$, $\lambda_F = 2 \cdot 10^{-6}$ and $\lambda_D = 0.01$.

Following the experimental results of Section 5.2, in each initialization, we consider $a_{i,j} = 0.1$ if it is considered as a non-existent edge in the provided structure and $i \neq j$ to avoid possible self-loops.

In practice, SAM is stable until it finds the moral graph, which we consider as the baseline of this graph. We consider as other possible initializations the true causal graph, the CPDAG, and the cases of adding, removing, or reversing an edge from the true causal graph. These last three cases will allow us to analyze SAM's stability when false knowledge is provided.

From Fig. 5.2 we can deduce that giving any information about the causal structure except if it contains a miss-oriented edge helps to improve its stability. The case with worst stability in all the cases was related to the graph structure $A \leftarrow B \rightarrow C$. A possible explanation of this issue can be the fact that both edges are oriented only based on conditional asymmetries. This result enforces the benefits of minimizing the structural loss to identify the true CPDAG.

However, from Fig. 5.2 we can conclude than removing an edge from the causal structure can result in better stability than giving the true causal graph. This result is strongly related with the experimental result of Section 5.2, where the structural gates are constant if fixed to zero. Indeed, a possible explanation can be that when the initial value of the coefficient $a_{i,j}$ is too low it might tend to go to zero due to the sparsity penalty on the loss. Hence, SAM might not be able to identify all the cases where false information is provided.

Chapter 6

Conclusion and future works

A causal framework provides a deeper comprehension of the relationships between variables and can accurately guide interventions. The high computational complexity of finding the causal graph based on observational data motivates the development of different approaches aiming to disentangle causes and effects. Among them, Structural Agnostic Model (SAM) [Kal+22] relies on a continuous-optimization based approach that uses a generative adversarial network (GAN). SAM presents some instability due to the architecture of neural networks.

This work studies the sources of SAM's instability and how to handle them.

Experimental results on simulated data showed that the initialization strategy comprises a major source of SAM's instability. Furthermore, we found that setting the initial probabilities of an edge being included in the DAG to zero help in partial handle it.

6.1 Future works

Several future work lines can be followed to improve our understanding of SAM's stability and enhance it.

Firstly, an interesting question might be to analyze the combined impact of the considered possible sources. These results might allow to design of new and more stable extensions of SAM taking into account the impact of these sources. For instance, by considering one or another way of increment the weight of the acyclicity constraint based on the initial provided structural gates.

Secondly, since it seems that fixing the $a_{i,j}$ coefficients to zero makes them remain constant, a possible way to improve SAM's stability can be to consider pruning edges during its training. This can result in a decrease in the weight of the acyclicity optimization method on the loss. Moreover, it might be interesting to analyze if there exists a trade-off between SAM's performance and its stability through this procedure.

Finally, the experimental results have shown that initializing SAM with information related to the causal structure might improve its stability. An exploration of this possibility can potentially lead to improved performance and stability. This motivates the question: could the performance be improved by considering an initialization obtained as a result of another causal discovery algorithm like the GES algorithm?

Bibliography

[BH77] William Thomas Bielby and Robert Mason Hauser. “Structural equation models”. In: *Annual review of sociology* 3.1 (1977), pp. 137–161.

[BP13] Kenneth A Bollen and Judea Pearl. “Eight myths about causality and structural equation models”. In: *Handbook of causal analysis for social research*. Springer, 2013, pp. 301–328.

[BE02] Olivier Bousquet and André Elisseeff. “Stability and generalization”. In: *The Journal of Machine Learning Research* 2 (2002), pp. 499–526.

[Chi96] David Maxwell Chickering. “Learning Bayesian networks is NP-complete”. In: *Learning from data: Artificial intelligence and statistics V* (1996), pp. 121–130.

[Chi02] David Maxwell Chickering. “Optimal structure identification with greedy search”. In: *Journal of machine learning research* 3.Nov (2002), pp. 507–554.

[CHM04] Max Chickering, David Heckerman, and Chris Meek. “Large-sample learning of Bayesian networks is NP-hard”. In: *Journal of Machine Learning Research* 5 (2004), pp. 1287–1330.

[CM+14] Diego Colombo, Marloes H Maathuis, et al. “Order-independent constraint-based causal structure learning.” In: *The Journal of Machine Learning Research* 15.1 (2014), pp. 3741–3782.

[Fan+16] Yi Fan et al. “Applications of structural equation modeling (SEM) in ecological studies: an updated review”. In: *Ecological Processes* 5 (2016), pp. 1–12.

[Goo+14] Ian Goodfellow et al. “Generative adversarial nets”. In: *Advances in neural information processing systems* 27 (2014).

[Gou+18] Olivier Goudet et al. “Learning functional causal models with generative neural networks”. In: *Explainable and interpretable models in computer vision and machine learning* (2018), pp. 39–80.

[HRS16] Moritz Hardt, Ben Recht, and Yoram Singer. “Train faster, generalize better: Stability of stochastic gradient descent”. In: *International Conference on Machine Learning*. PMLR. 2016, pp. 1225–1234.

[Hol86] Paul W Holland. “Statistics and causal inference”. In: *Journal of the American statistical Association* 81.396 (1986), pp. 945–960.

[IR15] Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

[KGD20] Diviyan Kalainathan, Olivier Goudet, and Ritik Dutta. “Causal discovery toolbox: Uncovering causal relationships in python”. In: *The Journal of Machine Learning Research* 21.1 (2020), pp. 1406–1410.

[Kal+22] Diviyan Kalainathan et al. “Structural agnostic modeling: Adversarial learning of causal graphs”. In: *The Journal of Machine Learning Research* 23.1 (2022), pp. 9831–9892.

[KB15] Diederik P. Kingma and Jimmy Ba. “Adam: A Method for Stochastic Optimization”. In: *3rd International Conference on Learning Representations*. Ed. by Yoshua Bengio and Yann LeCun. 2015.

[Kme20] Boris Kment. “Counterfactuals and causal reasoning”. In: *Perspectives on Causation: Selected Papers from the Jerusalem 2017 Workshop*. Springer. 2020, pp. 463–482.

[Lew73] David Lewis. “Causation”. In: *The journal of philosophy* 70.17 (1973), pp. 556–567.

[MMT16] Chris J. Maddison, A. Mnih, and Y. Teh. “The Concrete Distribution: A Continuous Relaxation of Discrete Random Variables”. In: *International Conference on Learning Representations*. 2016.

[AH10] Awad H Al-Mohy and Nicholas J Higham. “A new scaling and squaring algorithm for the matrix exponential”. In: *SIAM Journal on Matrix Analysis and Applications* 31.3 (2010), pp. 970–989.

[NGZ20] Ignavier Ng, AmirEmad Ghassami, and Kun Zhang. “On the role of sparsity and dag constraints for learning linear dags”. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 17943–17954.

[PBG17] Gordon B Parker, Heather Brotchie, and Rebecca K Graham. “Vitamin D and depression”. In: *Journal of affective disorders* 208 (2017), pp. 56–61.

[Pea09] Judea Pearl. *Causality: Models, Reasoning and Inference*. 2nd. Cambridge University Press, 2009.

[PV91] Judea Pearl and T Verma. *A theory of inferred causation. Principles of Knowledge Representation and Reasoning: Proceedings of the Second International Conference.* 1991.

[Pop22] Elena Popa. “Getting counterfactuals right: the perspective of the causal reasoner”. In: *Synthese* 200.1 (2022), p. 17.

[Ric+16] Nicole Franziska Richter et al. “A critical look at the use of SEM in international business research”. In: *International marketing review* 33.3 (2016), pp. 376–404.

[Sha48] C Shanon. *A Mathematical Theory of Communication. Bell System Technnical Journal, 27 (4), 623-656.* 1948.

[SGS00] Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, prediction, and search.* MIT press, 2000.

[VCB22] Matthew J Vowels, Necati Cihan Camgoz, and Richard Bowden. “D’ya like dags? a survey on structure learning and causal discovery”. In: *ACM Computing Surveys* 55.4 (2022), pp. 1–36.

[Woo05] James Woodward. *Making things happen: A theory of causal explanation.* Oxford University Press, 2005.

[Zhe+18] Xun Zheng et al. “Dags with no tears: Continuous optimization for structure learning”. In: *Advances in neural information processing systems* 31 (2018).