Structural Agnostic Modeling: Adversarial Learning of Causal Graphs

Diviyan Kalainathan*

DIVIYAN@LRI.FR

TAU, LRI, INRIA, CNRS, Université Paris-Saclay 660 Rue Noetzlin, Gif-Sur-Yvette, France

Olivier Goudet*

OLIVIER.GOUDET@UNIV-ANGERS.FR

LERIA, Université d'Angers 2 boulevard Lavoisier, 49045 Angers, France

Isabelle Guyon

GUYON@CHALEARN.ORG

TAU, LRI, INRIA, CNRS, Université Paris-Saclay 660 Rue Noetzlin, Gif-Sur-Yvette, France

David Lopez-Paz

DLP@FB.COM

Facebook AI Research 6 Rue Ménars, 75002 Paris

Michèle Sebag

SEBAG@LRI.FR

TAU, LRI, INRIA, CNRS, Université Paris-Saclay 660 Rue Noetzlin, Gif-Sur-Yvette, France

Editor:

Abstract

A new causal discovery method, *Structural Agnostic Modeling* (SAM), is presented in this paper. Leveraging both conditional independencies and distributional asymmetries in the data, SAM aims at recovering full causal models from continuous observational data along a multivariate non-parametric setting. The approach is based on a game between *d* players estimating each variable distribution conditionally to the others as a neural net, and an adversary aimed at discriminating the overall joint conditional distribution, and that of the original data. An original learning criterion combining distribution estimation, sparsity and acyclicity constraints is used to enforce the end-to-end optimization of the graph structure and parameters through stochastic gradient descent. Besides the theoretical analysis of the approach in the large sample limit, SAM is extensively experimentally validated on synthetic and real data.

1. Introduction

This paper addresses the problem of uncovering causal structure from multivariate observational data. This problem is receiving more and more attention with the increasing emphasis on model interpretability and fairness (Doshi-Velez and Kim, 2017).

While the gold standard to establish causal relationships remains randomized controlled experiments (Pearl, 2003a; Imbens and Rubin, 2015), in practice these often happen to be costly, unethical, or simply infeasible. Therefore, hypothesizing causal relations from observational data,

^{*} Equal contribution. This work was done during Olivier Goudet's post-doc at Univ. Paris-Saclay.

often referred to as *observational causal discovery*, has attracted much attention from the machine learning community (Lopez-Paz et al., 2015; Mooij et al., 2016; Peters et al., 2017). Observational causal discovery has found many applications, e.g. in economics to understand and model the impact of monetary policies (Chen et al., 2007), or in bio-informatics to infer network structures from gene expression data (Sachs et al., 2005) and prioritize confirmatory or exploratory experiments.

Observational causal discovery aims to learn both the causal graph and the associated causal mechanisms from samples of the joint probability distribution of observational data. Four

main approaches have been proposed in the literature (more in Section 2.4). A first approach exploits the Markov properties of Directed Acyclic Graphs (DAGs) in order to recover the Markov equivalence class of the causal graph, with the limitation that edges can be oriented only when detecting structural patterns, such as *v-structures*, and using specific propagation rules.

A second approach simultaneously learns the causal mechanisms and the causal graph structure with automatic regularization techniques to avoid a combinatorial search in the space of DAG. A third approach goes beyond the Markov equivalence class limitation by exploiting asymmetries in the joint distribution, based on the assumption that p(x)p(y|x) is simpler than p(y)p(x|y) (for some appropriate notion of simplicity) when X causes Y ($X \rightarrow Y$). A fourth approach combines methods from the first and third approaches. Another stream of work, closely related to causal discovery, is causal feature selection, which aims at recovering the Markov Blanket of given variables (Yu et al., 2018), extensively relying on estimating mutual information among variables (Bell and Wang, 2000; Brown et al., 2012; Vergara and Estévez, 2014).

Pertaining to the fourth approach of causal modeling, the contribution of this paper is a new causal discovery algorithm called *Structural Agnostic Modeling* (SAM), exploiting conditional independence relations *and* distributional asymmetries from observational continuous data. SAM relies on the general Functional Causal Model (FCM) framework (Pearl, 2003b), and makes no restriction on the complexity of the underlying causal mechanisms and data distributions.¹

SAM proceeds as follows: i) each causal mechanism in the FCM is a neural net trained from available data; ii) the combinatorial optimization problem, at the root of directed acyclic graph learning, is handled through sparsity and acyclicity constraints inspired from Leray and Gallinari (1999) and Yu et al. (2018); iii) the joint training of all causal mechanisms is handled through an adversarial approach (Goodfellow et al., 2014; Mirza and Osindero, 2014), enforcing the accuracy of the FCM joint distribution with respect to the data distribution. SAM also relies on Occam's razor principle to infer the causal graph, where the complexity of each candidate graph is evaluated using the Minimum Description Length (MDL). This causal inference principle will be assessed both theoretically and experimentally.

This paper is organized as follows: Section 2 introduces the problem of learning an FCM, presents the main underlying assumptions and briefly describes the state of the art in causal modeling. In section 3, an information theoretic approach is proposed to infer a causal graph. Section 4 describes the SAM algorithm devised to tackle the associated optimization problem and section 5 is devoted to the theoretical analysis of the approach. Section 6 presents the goals of experiments and the experimental setting used for the empirical validation of SAM. Section 7 reports on SAM empirical results compared to the state of the art. Section 8 discusses the contribution and presents some perspectives for future work.

^{1.} The SAM code is available at https://github.com/Diviyan-Kalainathan/SAM.

2. Observational Causal modeling: Formal Background

Let $\mathbf{X} = [X_1, \dots X_d]$ denote a vector of d continuous random variables, with unknown joint probability distribution $p(\mathbf{x})$. The observational causal discovery setting considers an iid n-sample drawn after $p(\mathbf{x})$, noted $D = \{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}\}$, with $\mathbf{x}^{(\ell)} = (x_1^{(\ell)}, \dots, x_d^{(\ell)})$ and $x_j^{(\ell)}$ the ℓ -th sample of X_j .

2.1 Functional Causal Models

The underlying generative model of the data is assumed to be a Functional Causal Model (FCM) (Pearl, 2003b), defined as a pair (\mathcal{G}, f) , with \mathcal{G} a directed acyclic graph and $f = (f_1, \dots, f_d)$ a set of d causal mechanisms. Formally, each variable X_i follows a distribution described as:

$$X_i \sim f_i(X_{\text{Pa}(i:\mathcal{G})}, E_i), \text{ with } E_i \sim \mathcal{N}(0, 1) \text{ for } j = 1, \dots, d.$$
 (1)

For notational simplicity, X_j denotes both a variable and the associated node in graph \mathcal{G} . Pa $(j;\mathcal{G})$ is the set of parents of X_j in \mathcal{G} , f_j is a function from $\mathbb{R}^{|\operatorname{Pa}(j;\mathcal{G})|+1} \to \mathbb{R}$ and E_j is a unit centered Gaussian noise², accounting for all unobserved causes of X_j .

A 5-variable FCM is depicted on Fig. 1.

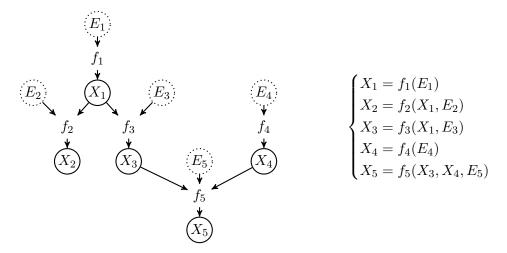


Figure 1: Example of a Functional Causal Model (FCM) on $\mathbf{X} = [X_1, \dots, X_5]$. Left: causal graph \mathcal{G} . Right: causal mechanisms.

2.2 Notations and Definitions

All notations used in the paper are listed in Appendix A.

 $\mathbf{X}_{\setminus i}$ denotes the set of all variables but X_i .

Conditional independence: $(X_i \perp \!\!\! \perp X_j | X_k)$ means that variables X_i and X_j are independent conditionally to X_k , i.e. $P(X_i, X_j | X_k) = P(X_i | X_k) P(X_j | X_k)$.

Markov blanket: a Markov blanket $MB(X_i)$ of a variable X_i is a minimal subset of variables in $X_{\setminus i}$

^{2.} Note that this is not an actual restriction on the FCM space, as any type of noise can be modeled as some g(E) with g a function and E a Gaussian noise (Stegle et al., 2010).

such that any disjoint set of variables in the network is independent of X_i conditioned on $MB(X_i)$. **V-structure**: Variables $\{X_i, X_j, X_k\}$ form a v-structure iff their causal structure is: $X_i \to X_k \leftarrow X_j$. **Skeleton of the DAG**: the skeleton of the DAG is the undirected graph obtained by replacing all edges by undirected edges.

Markov equivalent DAG: two DAGs with same skeleton and same v-structures are said to be *Markov equivalent* (Pearl and Verma, 1991). A *Markov equivalence class* is represented by a *Completed Partially Directed Acyclic Graph* (CPDAG) having both directed and undirected edges.

Variables X_i and X_j are said to be adjacent according to a CPDAG iff there exists an edge between both nodes. If directed, this edge models causal relationship $X_i \to X_j$ or $X_j \to X_i$. If undirected, it models a causal relationship in either direction.

2.3 Causal Assumptions and Properties

In this paper, the recovery of the underlying causal graph \mathcal{G} from observational data relies on the following assumptions:

Acyclicity: The causal graph \mathcal{G} (Eq. (1)) is assumed to be a Directed Acyclic Graph (DAG).

Causal Markov Assumption (CMA): Noise variables E_j (Eq. (1)) are assumed to be independent from each other. This assumption together with the above DAG assumption yields the classical causal Markov property, stating that all variables are independent of their non-effects (non descendants in the causal graph) conditionally to their direct causes (parents) (Spirtes et al., 2000). Under the causal Markov assumption, the distribution described by the FCM satisfies all conditional independence relations³ among variables in \mathbf{X} via the notion of d-separation (Pearl, 2009). Accordingly the joint distribution $p(\mathbf{x})$ can be factorized as the product of the distributions of each variable conditionally on their parents in the graph:

$$p(\mathbf{x}) = \prod_{j=1}^{d} p(x_j | x_{\text{Pa}(j;\mathcal{G})})$$
(2)

Causal Faithfulness Assumption (CFA): The joint distribution $p(\mathbf{x})$ is assumed to be *faithful* to graph \mathcal{G} , that is, every conditional independence relation that holds true according to p is entailed by \mathcal{G} (Spirtes and Zhang, 2016). It follows from causal Markov and faithfulness assumptions that every causal path in the graph corresponds to a dependency between variables, and vice versa.

Causal Sufficiency assumption (CSA): X is assumed to be *causally sufficient*, that is, a pair of variables $\{X_i, X_j\}$ in X has no common cause external to $X_{\setminus i,j}$.

2.4 Background

This section briefly presents a formal background of observational causal discovery, referring the reader to (Spirtes et al., 2000; Peters et al., 2017) for a comprehensive survey.

Observational causal discovery algorithms are structured along four categories:

I The first category aims to recover the Markov equivalence class of the DAG using conditional independencies. One option is based on backward selection, starting from a complete graph

^{3.} It must be noted however that the data might satisfy additional independence relations beyond those in the graph; see the faithfulness assumption.

and removing edges based on conditional independence tests (Spirtes et al., 2000). Another option associates a score with candidate causal graph, and performs a combinatorial search to find the best candidate according to that score (Chickering, 2002).

In all these approaches, the doubly exponential-size DAG search space is explored using local search, thus facing severe scalability issues. Although many heuristics have been deployed to explore the DAG space (see e.g., (Tsamardinos et al., 2006)), these remain impractical for a high number of variables.

- II A second category of approaches relies on assumptions on the underlying generative process to avoid the combinatorial exploration of the DAG search space. For instance, assuming Gaussian data distribution, Meinshausen and Bühlmann (2006) recovers the causal DAG skeleton by examining the non-zero entries in the inverse covariance matrix of the data. Along the same line, Shojaie and Michailidis (2010); Ren et al. (2016) assume sparse causal graphs and leverage Lasso-type regression techniques to extract the strongest causal relations. The underlying assumptions (linear causal mechanisms with additive Gaussian noise) however entail severe restrictions. On the one hand, real-world data is not always Gaussian; on the other hand, linear-Gaussian causal mechanisms do not involve any asymmetries that could be leveraged as causal footprints in the data.⁴
- III The third category of approaches exploits such asymmetries or causal footprints in the data generative process to uniquely identify the causal DAG. According to Quinn et al. (2011), the first approach in this direction is LiNGAM (Shimizu et al., 2006). LiNGAM handles linear structural equation models on continuous variables, where each variable is modeled as the weighted sum of its parents and noise. Assuming further that all noise variables are non-Gaussian, Shimizu et al. (2006) show that the causal structure is fully identifiable (all edges can be oriented).

This category of methods has mainly be applied in the continuous, non-linear bivariate case where conditional independence tests cannot be used to uncover the causal relation, e.g. the Additive Noise Model (Hoyer et al., 2009), the Gaussian Process Inference causal model (Stegle et al., 2010), and the Randomized Causation Coefficient (Lopez-Paz et al., 2015). A main merit of such bivariate methods is to independently orient each edge (with no propagation and thus no risk of error propagation). In counterpart, bivariate methods do not have a global view of the variable set, and specifically cannot take advantage of v-structures. For instance when considering the v-structure $X \to Z \leftarrow Y$, a bivariate model based on cause-effect asymmetry would miss both causal relations in the case of Gaussian distributions of variables and noise, and linear mechanism.

IV The fourth category — including the Causal Additive Model (Bühlmann et al., 2014) and the Causal Generative Neural Networks (Goudet et al., 2018) — tackles the multivariate causal discovery problem by leveraging conditional independence relations and distributional asymmetries in the data. Both approaches however suffer from the same scalability limitations as the first category of approaches, as they face the exploration of the DAG search space.

^{4.} Typically, in domains such as biology, the sought \mathcal{G} graph is star-shaped and does not include v-structures. In such cases, the approaches based on Gaussianity assumptions are unable to orient the edges (see section 7.2).

The proposed SAM approach ambitions to combine the best of all the above: exploiting conditional independence relations as all approaches but those in category III; using regularization terms to avoid combinatorial optimization, like approaches in category II; and exploiting distributional asymmetries, like categories III and IV.

3. Causal modeling through data compression

This section describes the information theory framework underlying the proposed SAM approach, first introducing the notion of Kolmogorov complexity (Li and Vitányi, 2013), and describing a computable approximation thereof in terms of Minimum Description Length (Grünwald, 2007).

3.1 Kolmogorov complexity

According to the above-mentioned assumptions in section 2.3, the generative model underlying the data is an acyclic FCM based on graph \mathcal{G} , and the joint distribution $p(\mathbf{x})$ of the observational data can be factored as

$$p(\mathbf{x}) = \prod_{j=1}^{d} p(x_j | x_{\text{Pa}(j;\mathcal{G})}).$$

Furthermore we assume that the ground truth model is the simplest model accounting for the data distribution (Occam's razor principle). Formally, the factorization of the joint distribution $p(\mathbf{x})$ along any other candidate DAG $\hat{\mathcal{G}} \neq \mathcal{G}$ is assumed to be more complex, in the sense defined below.

It is emphasized that the Occam's razor principle goes beyond classical Bayesian approaches as it supports the ranking of Markov-equivalent solutions. This principle is at the core of the cause effect pair approaches: as both DAGs $X \to Y$ and $Y \to X$ are Markov equivalent, one further assumes that the factorization of p(x,y) into p(x)p(y|x) is of lower complexity than the alternative factorization p(y)p(x|y) if the *true* DAG is $X \to Y$ (Stegle et al., 2010). In the general multi-variate setting, irrespective of the model search space, the Occam's razor principle has been formalized by Janzing and Scholkopf (2010) in terms of Kolmogorov complexity.

Referring the reader to Li and Vitányi (2013) for a comprehensive introduction, the Kolmogorov complexity of a probability distribution p of the continuous variable X defined on dom(X) is the description length of the shortest program that implements its sampling process (Grünwald et al., 2008) (Eq. 14), noted K(p) (also noted $K(p(\mathbf{x}))$) in the following by abuse of notation):

$$K(p) = \min_{s} \left\{ |s| \ : \ \text{for all} \ m \ \in \{1, 2, \ldots\}, x \in dom(X) : |\mathbb{U}(s, x, m) - p(x)| \le 1/m \right\}, \quad (3)$$

with \mathbb{U} a Universal Turing machine. Taking inspiration from (Janzing and Scholkopf, 2010), the key working hypothesis in the remainder of the paper is that the sought causal models are those with minimum Kolmogorov complexity of their conditional probabilities:

Working Hypothesis 1 (Algorithmic independence of statistical properties) (Janzing and Scholkopf, 2010)

A necessary condition for causal model G (i.e., a DAG) to hold is that the shortest description of the joint density p be the sum of the shortest description of its causal mechanisms, up to a constant:

$$K(p(\mathbf{x})) \stackrel{+}{=} \sum_{j=1}^{d} K(p(x_j|x_{Pa(j;\mathcal{G})})), \tag{4}$$

where $\stackrel{+}{=}$ denotes equality up to an additive constant.

The right hand side in Eq. (4) is the overall Kolmogorov complexity of causal model \mathcal{G} , which must equal the Kolmogorov complexity of the whole joint distribution p to be admissible.

3.2 Minimum Description Length

As the Kolmogorov complexity is not computable however, a tractable approximation thereof, the Minimum Description Length (MDL) is often used in practice, in particular in relation with bivariate causal discovery (Stegle et al., 2010; Budhathoki and Vreeken, 2017). This theoretical framework is extended to the multivariate setting as follows.

Let p be defined after causal graph $\widehat{\mathcal{G}}$. The MDL associated with p measured with respect to a class \mathcal{Q} of computable probabilistic models (e.g. exponential models), and an iid n-sample drawn after $p(\mathbf{x})$ noted $D = \{\mathbf{x}^{(1)} \dots \mathbf{x}^{(n)}\}$ is defined as (Barron and Cover, 1991):

$$MDL_r(\widehat{\mathcal{G}}, D) := \min_{q \text{ in } \mathcal{Q}} \left[K(q(\mathbf{x}, \widehat{\mathcal{G}})) + \sum_{\ell=1}^n \log \frac{1}{q(\mathbf{x}^{(\ell)}, \widehat{\mathcal{G}})} \right]$$
 (5)

with $K(q(\mathbf{x},\widehat{\mathcal{G}}))$ the number of bits needed to describe model q (that is computable by definition of \mathcal{Q}) and $\sum_{\ell=1}^n \log \frac{1}{q(\mathbf{x}^{(\ell)},\widehat{\mathcal{G}})}$ the number of bits in the coding length of the dataset with respect to q.

The MDL used in the following is the normalized MDL, divided by the size n of the iid-sample D:

$$MDL(\widehat{\mathcal{G}}, D) := \min_{q \text{ in } \mathcal{Q}} \left[\frac{1}{n} K(q(\mathbf{x}, \widehat{\mathcal{G}})) + \frac{1}{n} \sum_{\ell=1}^{n} \log \frac{1}{q(\mathbf{x}^{(\ell)}, \widehat{\mathcal{G}})} \right]$$
(6)

Causal inference with Minimum Description Length Overall, the working hypothesis is that the Kolmogorov complexity of the true \mathcal{G} , and the MDL-based approximation MDL(\mathcal{G} , D) thereof, are minimal. If the minimal MDL is reached for a *unique* DAG \mathcal{G}^* , this graph is therefore the sought causal model under the assumptions made. Note however that the unicity of the solution is not guaranteed.

A well-known example is the linear bivariate Gaussian model, with Y = X + E and $X \perp \!\!\! \perp E$, with X and E Gaussian variables. As established by Mooij et al. (2016), there exists two models q_1 and q_2 such that $p(x) = q_1(x)q_1(y|x) = q_2(y)q_2(x|y)$ with exact same *complexity* (same structure and same number of parameters). In such cases, $MDL(X \to Y, D)$ and $MDL(Y \to X, D)$ are equal in the large sample limit and the causal graph remains undetermined.

4. Structural Agnostic model

This section presents the *Structural Agnostic Model* (SAM), implementing the MDL framework presented in the last section within the space of generative neural networks (NN). The originality of

the approach is to implement an end-to-end search for a Functional Causal Model (FCM, Eq. (1)) with **no restrictive assumption on the underlying causal mechanisms and data distributions**.

4.1 Modeling causal mechanisms with conditional generative neural networks

The model search space includes all distributions q defined from a DAG $\widehat{\mathcal{G}}$ and causal mechanisms $\widehat{f} = (\widehat{f}_1, \dots, \widehat{f}_d)$, with \widehat{f}_j a 1-hidden layer NN yielding a generative model of X_j from all other variables in \mathbf{X} (Fig. 2). Formally:

- The d-dimensional vector of variables \mathbf{X} is elementwise multiplied with binary vector $\mathbf{a}_j = (a_{1,j}, \dots a_{d,j})$ named structural gate. Coefficient $a_{i,j}$ is 1 iff variable X_i is used to generate X_j (with $a_{i,i}$ set to 0 to avoid self-loops), that is, edge $X_i \to X_j$ is present in graph $\widehat{\mathcal{G}}$, and X_i is considered to be a cause of X_j . Otherwise, $a_{i,j}$ is set to 0. A regularization term on \mathbf{a}_j enforces the graph sparsity.
- The number of active hidden units in neural network \hat{f}_j is controlled by a Boolean vector \mathbf{z}_j of size n_h named functional gate, where the h-th entry noted $z_{h,j} \in \{0,1\}$ corresponds to the activation of the h-th hidden unit of the neural network. Likewise, a regularization on the functional gates is used to limit the complexity of the functional mechanisms.
- At every evaluation of noise variable E_j , a value is drawn anew from distribution $\mathcal{N}(0,1)$. As already mentioned (footnote 1) the restriction to Gaussian noise is not a limitation.

As said, \hat{f}_j is implemented as a 1-hidden layer NN, i.e. a linear combination of non-linear features $\phi_{i,k}$:

$$X_{j} = \hat{f}_{j}(\mathbf{X}, E_{j}) = \sum_{k=1}^{n_{h}} m_{j,k} \phi_{j,k}(\mathbf{X}, E_{j}) z_{j,k} + m_{j,0}$$
with $\phi_{j,k}(\mathbf{X}, E_{j}) = tanh\left(\sum_{i=1}^{d} W_{j,i} a_{j} X_{j} + W_{j,0} + W_{j,d+1} E_{j}\right)$
(7)

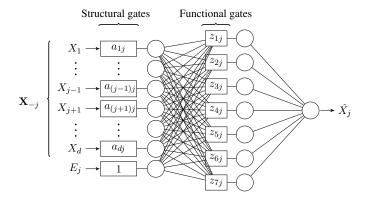


Figure 2: Diagram of the conditional generative neural network modeling the causal mechanism $X_j = \hat{f}_j(\mathbf{X}, E_j)$.

For notational simplicity, each \hat{f}_j is associated with a parameter vector $\theta_j = (\theta_{j,1}, \dots, \theta_{j,p_j})$ (including vectors \mathbf{m}_j and $W_{j,\cdot}$ but excluding the \mathbf{a}_j and \mathbf{z}_j gates).

With E_j a Gaussian noise variable, each \hat{f}_j thus encodes a generative model of X_j conditionally to variables in $x_{\text{Pa}(j;\widehat{\mathcal{G}})}$, with $\text{Pa}(j;\widehat{\mathcal{G}}) = \{i \in [1,\ldots,d] \ s.t. \ a_{i,j} = 1\}$.

Under the assumptions that noise variables E_j are independent of each other, and graph $\widehat{\mathcal{G}}$ is acyclic, noting θ the concatenation of parameters θ_1,\ldots,θ_d and $Z=\{z_{h,j}\}$ the functional gate $n_h\times d$ matrix, the candidate model $(\widehat{\mathcal{G}},\widehat{f})$ defines a multivariate distribution $q(\mathbf{x},\widehat{\mathcal{G}},\theta,Z)$ after the global Markov property:

$$q(\mathbf{x}, \widehat{\mathcal{G}}, \theta, Z) = \prod_{j=1}^{d} q(x_j | x_{\text{Pa}(j; \widehat{\mathcal{G}})}, \theta_j, \mathbf{z}_j)$$
(8)

Moreover, as the conditional densities $q(x_j|x_{\text{Pa}(j;\widehat{\mathcal{G}})},\theta_j,\mathbf{z}_j)$ can be computed independently,

$$K(q(\mathbf{x},\widehat{\mathcal{G}},\theta,Z)) \stackrel{+}{=} \sum_{j=1}^{d} K(q(x_j|x_{\operatorname{Pa}(j;\widehat{\mathcal{G}})},\theta_j,\mathbf{z}_j)).$$

The normalized MDL for a candidate graph $\widehat{\mathcal{G}}$ (Eq. (5)) thus is rewritten as a sum of d local scores:

$$MDL(\widehat{\mathcal{G}}, \theta^*, D) = \min_{\theta, Z} \left[\underbrace{\frac{1}{n} \sum_{j=1}^{d} K(q(x_j | x_{\text{Pa}(j; \widehat{\mathcal{G}})}, \theta_j, \mathbf{z}_j))}_{\text{model complexity}} + \underbrace{\frac{1}{n} \sum_{j=1}^{d} \sum_{\ell=1}^{n} \log \frac{1}{q(x_j^{(\ell)} | x_{\text{Pa}(j; \widehat{\mathcal{G}})}^{(\ell)}, \theta_j, \mathbf{z}_j)}}_{\text{fit loss}} \right]$$

with θ^* the optimal set of parameters for the considered model.

4.2 SAM learning criterion

This section derives a principled loss function from the model complexity and data fitting terms in Eq. (9), defining SAM learning criterion.

Model complexity While
$$K(q(x_j|x_{\text{Pa}(j:\widehat{\mathcal{G}})}, \theta_j, \mathbf{z}_j))$$

could be estimated using the Akaike Information or the Bayesian Information Criterion, the complexity of the graph structure and of the causal mechanisms can by construction be assessed and controlled through respectively the L_0 norm of the structural and functional gates \mathbf{a}_j and \mathbf{z}_j (that is, the number of parents of X_j and the number of effective neurons in \hat{f}_j):

$$K(q(x_j|x_{\operatorname{Pa}(j;\widehat{\mathcal{G}})}, \theta_j, \mathbf{z}_j)) \stackrel{\text{def}}{=} \lambda_S |\operatorname{Pa}(j;\widehat{\mathcal{G}})| + \lambda_F \sum_{h=1}^{n_h} z_{h,j}$$
(10)

with $\lambda_S>0$ and $\lambda_F>0$ the regularization weights. For notational simplicity we write $q(x_j|x_{\mathrm{Pa}(j;\widehat{\mathcal{G}})},\theta_j)$ instead of $q(x_j|x_{\mathrm{Pa}(j;\widehat{\mathcal{G}})},\theta_j,\mathbf{z}_j)$ in the following.

Data fitting loss As said, when the number of samples $\mathbf{x}^{(\ell)}$ goes to infinity, the data fitting term goes to data log-likelihood expectation under the sought generative distribution:

$$\lim_{n \to \infty} \frac{1}{n} \sum_{\ell=1}^{n} \log \frac{1}{q(x_j^{(\ell)} | x_{\mathsf{Pa}(j;\widehat{\mathcal{G}})}^{(\ell)}, \theta_j)} = -\mathbb{E}_p \log q(x_j | x_{\mathsf{Pa}(j;\widehat{\mathcal{G}})}, \theta_j)$$
(11)

For $j=1\ldots d$, for $\mathbf{x}=(x_1,\ldots,x_d)$ let \mathbf{x}_{-j} be defined as $(x_1,\ldots x_{j-1},x_{j+1},\ldots,x_d)$. The distribution of x_j conditionally to \mathbf{x}_{-j} is denoted as $q(x_j|\mathbf{x}_{-j})$. Considering FCM $(\widehat{G},\widehat{f})$, as variable X_j only depends on $X_{\mathrm{Pa}(j:\widehat{G})}$, it follows that $q(x_j|\mathbf{x}_{\mathrm{Pa}(j:\widehat{G})},\theta_j)=q(x_j|\mathbf{x}_{-j},\theta_j)$. Therefore:

$$\lim_{n \to \infty} \frac{1}{n} \sum_{\ell=1}^{n} \log \frac{1}{q(x_j^{(\ell)} | \mathbf{x}_{\mathbf{p}_0(j;\widehat{G})}^{(\ell)}, \theta_j)} = \mathbb{E}_p \log \frac{1}{q(x_j | \mathbf{x}_{-j}, \theta_j)}$$
(12)

$$= \mathbb{E}_p \log \frac{p(x_j|\mathbf{x}_{-j})}{q(x_j|\mathbf{x}_{-j},\theta_j)} - \mathbb{E}_p \log p(x_j|\mathbf{x}_{-j})$$
(13)

$$= D_{KL}[p(x_j|\mathbf{x}_{-j}) || q(x_j|\mathbf{x}_{-j}, \theta_j)] + H(X_j|\mathbf{X}_{-j}), \quad (14)$$

with $D_{KL}[p(x_j|\mathbf{x}_{-j}) \parallel q(x_j|\mathbf{x}_{-j},\theta_j)]$ the Kullback-Leibler divergence between the true conditional distribution $p(x_j|\mathbf{x}_{-j})$ and $q(x_j|\mathbf{x}_{-j},\theta_j)$, and $H(X_j|\mathbf{X}_{-j})$ the constant, domain-dependent entropy of X_j conditionally to \mathbf{X}_{-j} (neglected in the following).

Taking inspiration from Nguyen et al. (2010); Nowozin et al. (2016), $D_{KL}[p(x_j|\mathbf{x}_{-j}) \parallel q(x_j|\mathbf{x}_{-j},\theta_j)]$ is estimated using an adversarial approach. Formally, for j=1 to d, for each initial sample $\mathbf{x}^{(\ell)}$ let pseudo-sample $\tilde{\mathbf{x}}_j^{(\ell)}$ be defined from $\mathbf{x}^{(\ell)}$ by replacing its j-th coordinate by $\hat{f}_j(\mathbf{x}^{(\ell)},e_j^{(\ell)})$, with $e_j^{(\ell)}$ drawn from $\mathcal{N}(0,1)$. Let dataset \tilde{D}_j denote the set of all pseudo $\tilde{\mathbf{x}}_j^{(\ell)}$ for $\ell=1$ to n. Let T_ω be a neural net trained to discriminate between the original dataset D on the one hand,

Let T_{ω} be a neural net trained to discriminate between the original dataset D on the one hand, and the dataset $\tilde{D} = \bigcup_{j=1}^{d} \tilde{D}_{j}$. Then, the scaled log-likelihood of the data in the large sample limit can be approximated after Nguyen et al. (2010):

$$\frac{1}{n} \sum_{j=1}^{d} \sum_{\ell=1}^{n} \log \frac{1}{q(x_{j}^{(\ell)} | x_{\text{Pa}(j;\widehat{\mathcal{G}})}^{(\ell)}, \theta_{j})} \approx \sup_{\omega \in \Omega} \left(\frac{d}{n} \sum_{\ell=1}^{n} T_{\omega}(\mathbf{x}^{(\ell)}) + \frac{1}{n} \sum_{j=1}^{d} \sum_{\ell=1}^{n} \left[-\exp(T_{\omega}(\tilde{\mathbf{x}}_{j}^{(\ell)}) - 1) \right] \right) + constant \quad (15)$$

$$D_{KL}[p(x_j, \mathbf{x}_{-j}) \parallel q(x_j, \mathbf{x}_{-j}, \theta_j)] \approx \sup_{\omega \in \Omega_j} \left(\frac{1}{n} \sum_{\ell=1}^n [T_\omega^j(\mathbf{x}^{(\ell)})] + \frac{1}{n} \sum_{\ell=1}^n [-\exp(T_\omega^j(\tilde{\mathbf{x}}_j^{(\ell)}) - 1)] \right)$$
(16)

Note that using a single discriminator T_{ω} to discriminate among D and \tilde{D} is more computationally efficient than building d discriminators (among D and each \tilde{D}_i) and yields a more stable algorithm.⁵

^{5.} It avoids the gradient vanishing phenomena that were empirically observed when building d discriminators.

Evaluation of the global loss min-max penalized optimization problem with SAM Overall, SAM is trained by solving a min-max penalized optimization problem (Eqs (10) for the model complexity and (15) for the data fitting term):

$$MDL(\widehat{\mathcal{G}}, \theta^*, D) = \min_{Z, A, \theta} \left(\underbrace{\frac{\lambda_S}{n} \sum_{i,j} a_{i,j} + \frac{\lambda_F}{n} \sum_{h,j} z_{h,j}}_{\text{model complexity}} + \underbrace{\frac{d}{n} \sum_{\omega \in \Omega} \left(\frac{d}{n} \sum_{\ell=1}^{n} T_{\omega}(\mathbf{x}^{(\ell)}) + \frac{1}{n} \sum_{j=1}^{d} \sum_{\ell=1}^{n} [-\exp(T_{\omega}(\tilde{\mathbf{x}}_{j}^{(\ell)}) - 1)] \right)}_{\text{fit loss}} \right), \quad (17)$$

where the minimization is carried over the set of parameters $\theta = (\theta_1, \dots, \theta_d)$ of the generators and over the matrices A and Z representing the structural and functional gates.

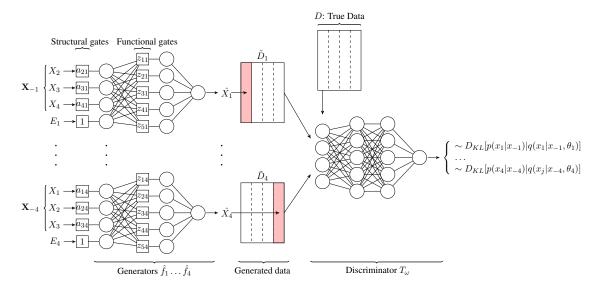


Figure 3: A four-variable example: Diagram of the SAM structure for variables X_1, \ldots, X_4

Fig. 3 illustrates a 4-variable SAM: on the left are the four generators corresponding to the causal mechanisms $\hat{f}_j^{\theta_j,a_j,z_j}$, for $j=1\dots 4$. On the right is the shared neural network discriminator T_ω evaluating the global fit loss corresponding to the sum of the estimated fit terms $D_{KL}[p(x_j,\mathbf{x}_{-j}) \mid\mid q(x_j,\mathbf{x}_{-j},\theta_j)]$ for $j=1\dots 4$.

4.3 Enforcing the acyclicity of the causal graph

Note that Eq. (17) does not ensure that the optimal $\widehat{\mathcal{G}}$ be a DAG: the sparsity constraint on $\widehat{\mathcal{G}}$ through the model complexity term (minimizing $\|\mathbf{a}_j\|_0$) leads to independently identify the Markov blanket of each variable X_j , selecting all causes, effects and spouses thereof (Yu et al., 2018).

In order to ensure that the solution is a DAG and avoid the associated combinatorial optimization issues (section 2.4),

it is proposed to augment the learning criterion with an acyclicity term inspired from Zheng et al. (2018b). The use of other acyclicity characterizing criteria Zheng et al. (2018a) is left for further work. Letting A denote the structural gate matrix (the adjacency matrix of the graph), $\widehat{\mathcal{G}}$ is a DAG iff

$$\sum_{k=1}^{d} \frac{\operatorname{tr} A^k}{k!} = 0$$

Accordingly, the learning criterion is augmented with an acyclicity term, with:

$$MDL(\widehat{\mathcal{G}}^*, \theta^*, D) = \min_{A, Z, \theta} \max_{\omega \in \Omega} \left(\frac{1}{n} \sum_{\ell=1}^{n} \sum_{j=1}^{d} [T_{\omega}(\mathbf{x}^{(\ell)}) - \exp(T_{\omega}(\widetilde{\mathbf{x}}_{j}^{(\ell)} - 1))] + \frac{\lambda_S}{n} \sum_{i,j} a_{i,j} + \frac{\lambda_F}{n} \sum_{j,h} z_{h,j} + \lambda_D \sum_{k=1}^{d} \frac{\operatorname{tr} A^k}{k!} \right),$$
(18)

with $\lambda_D > 0$ a penalization weight.⁶

This acyclicity constraint creates a coupling among the d feature selection problems, implying that at most one arrow between pairs of variables can be selected, and more generally leading to remove effect variables from the set of parents of any X_i ; the removal of effect variables in turn leads to removing spouse variables as well (section 5.1).

As the use of the L_0 norms of **as** and **zs**, if naively done, could entail computational issues (retraining the network from scratch for every new graph structure or neural architecture), an approach based on the Bernoulli reparameterization trick is proposed to end-to-end train the SAM architecture and weights using stochastic gradient descent (Srivastava et al., 2014; Louizos et al., 2017) and the Binary Concrete relaxation approach (Maddison et al., 2016; Jang et al., 2016). This solution corresponds to a learned dropout of edges and hidden units of the neural network.

Overall, the optimization of the learning criterion in Eq.(18) with the acyclicity and sparsity constraints defines the *Structural Agnostic Model* SAM (Alg. 1, Fig. 3).

5. Theoretical Analysis

This section analyzes the MDL learning criterion, decomposed into two terms: a structural loss and a parametric loss. It is finally shown that under some mild assumptions SAM recovers the true underlying graph \mathcal{G} .

Using Eq. (10), Eq. (9) can be rewritten as:

$$MDL(\widehat{\mathcal{G}}, \theta^*, D) = \frac{\lambda_S}{n} |\widehat{\mathcal{G}}| + \frac{\lambda_F}{n} \sum_{j=1}^d \|\mathbf{z}_j\|_0 + \frac{1}{n} \sum_{j=1}^d \sum_{\ell=1}^n \log \frac{1}{q(x_j^{(\ell)} | x_{\text{Pa}(j;\widehat{\mathcal{G}})}^{(\ell)}, \theta_j^*)}$$
(19)

According to (Brown et al., 2012), each scaled conditional log-likelihood term can be decomposed into three terms as:

^{6.} In practice, λ_D is small at the initialization and increases along time; in this way, the structural penalization term $\lambda_S \sum_{i,j} a_{i,j}$ can operate and prune the less relevant edges before considering the DAG constraint.

Algorithm 1 The Structural Agnostic Modeling Algorithm

for number of iterations do

for j = 1, ..., d **do**

- sample the structural gate vector \mathbf{a}_j : for $i=1,\ldots,d$, $a_{i,j}=\operatorname{cst}(H(l_{i,j}+a'_{i,j}))-\operatorname{cst}(\operatorname{sigmoid}(l_{i,j}+a'_{i,j}))+\operatorname{sigmoid}(l_{i,j}+a'_{i,j})$ with $l_{i,j}$ drawn from logistic distribution and H the Heavyside step function a.
- sample the functional gate vector \mathbf{z}_j : for $h=1,\ldots,n_h,\,z_{h,j}=\operatorname{cst}(H(l_{h,j}+z'_{h,j}))-\operatorname{cst}(\operatorname{sigmoid}(l_{h,j}+z'_{h,j}))+\operatorname{sigmoid}(l_{h,j}+z'_{h,j})$ with $l_{h,j}$ drawn from logistic distribution.
- sample noise variables, $e_i^{(\ell)} \sim \mathcal{N}(0,1)$ for $\ell = 1 \dots n, j = 1 \dots d$.
- \bullet generate n samples $\{\tilde{\mathbf{x}}_j^{(\ell)}\}_{l=1}^n$ such that for $\ell=1\dots,n$:

$$\begin{split} \tilde{x}_{j}^{(\ell)} &= \hat{f}_{j}^{\theta_{j}, a_{j}, z_{j}}(\mathbf{x}_{-j}^{(\ell)}, e_{j}^{(\ell)}) \\ &= \sum_{k=1}^{n_{h}} m_{j, k} tanh\left(\sum_{i=1}^{d} W_{j, i} a_{j} X_{j} + W_{j, 0} + W_{j, d+1} E_{j}\right) + m_{j, 0} \end{split}$$

end for

• update the discriminator by ascending its stochastic gradient:

$$\nabla_{\omega} \left[\frac{d}{n} \sum_{\ell=1}^{n} T_{\omega}(\mathbf{x}^{(\ell)}) + \frac{1}{n} \sum_{j=1}^{d} \sum_{\ell=1}^{n} \left[-\exp(T_{\omega}(\tilde{x}_{j}^{(\ell)}, \mathbf{x}_{-j}^{(\ell)}) - 1) \right] \right]$$

for j = 1, ..., d **do**

• update the generator by descending its stochastic gradients w.r.t the set of parameters $\theta_j = (m_j, W_j, n_j, b_j, \beta_j)$, the set of parameters \mathbf{a}'_j of the structural gates \mathbf{a}_j and the set of parameters \mathbf{z}'_j of the functional gates \mathbf{z}_j :

$$\begin{split} \nabla_{j} = & \nabla_{\theta_{j}} \left[\frac{1}{n} \sum_{\ell=1}^{n} \left[-\exp(T_{\omega}(\tilde{x}_{j}^{(\ell)}, \mathbf{x}_{-j}^{(\ell)}) - 1) \right] \right] \\ & + \nabla_{\mathbf{a}_{j}'} \left[\frac{1}{n} \sum_{\ell=1}^{n} \left[-\exp(T_{\omega}(\tilde{x}_{j}^{(\ell)}, \mathbf{x}_{-j}^{(\ell)}) - 1) \right] + \frac{\lambda_{S}}{n} \sum_{i,j} a_{i,j} + \lambda_{D} \sum_{k=1}^{d} \frac{\operatorname{tr} A^{k}}{k!} \right] \\ & + \nabla_{\mathbf{z}_{j}'} \left[\frac{1}{n} \sum_{\ell=1}^{n} \left[-\exp(T_{\omega}(\tilde{x}_{j}^{(\ell)}, \mathbf{x}_{-j}^{(\ell)}) - 1) \right] + \frac{\lambda_{F}}{n} \sum_{j,h} z_{h,j} \right] \end{split}$$

end for

end for

Return A and $\hat{f}_1, \ldots, \hat{f}_d$

a. cst() represents the copy by value operator transforming the input into a constant with the same value but zero gradient. With this trick the value of $a_{i,j}$ is equal to $H(l_{i,j} + a'_{i,j})$ (forward pass) but its gradient w.r.t $a'_{i,j}$ is equal to $\nabla_{a'_{i,j}}$ sigmoid $(l_{i,j} + a'_{i,j})$ (backward pass).

$$\frac{1}{n} \sum_{\ell=1}^{n} \log \frac{1}{q(x_{j}^{(\ell)} | x_{\text{Pa}(j;\widehat{\mathcal{G}})}^{(\ell)}, \theta_{j}^{*})} = \frac{1}{n} \sum_{\ell=1}^{n} \log \frac{p(x_{j}^{(\ell)} | x_{\text{Pa}(j;\widehat{\mathcal{G}})}^{(\ell)})}{q(x_{j}^{(\ell)} | x_{\text{Pa}(j;\widehat{\mathcal{G}})}^{(\ell)}, \theta_{j}^{*})} + \frac{1}{n} \sum_{\ell=1}^{n} \log \frac{p(x_{j}^{(\ell)} | \mathbf{x}_{-j}^{(\ell)})}{p(x_{j}^{(\ell)} | x_{\text{Pa}(j;\widehat{\mathcal{G}})}^{(\ell)})} + \frac{1}{n} \sum_{\ell=1}^{n} \log \frac{p(x_{j}^{(\ell)} | \mathbf{x}_{-j}^{(\ell)})}{p(x_{j}^{(\ell)} | \mathbf{x}_{-j}^{(\ell)})}$$

$$(20)$$

Note that term $\frac{1}{n}\sum_{\ell=1}^n \log p(x_j^{(\ell)}|\mathbf{x}_{-j}^{(\ell)})$ is a domain-dependent constant, converging toward $H(X_j|\mathbf{X}_{-j})$, the negative entropy of X_j conditionally to \mathbf{X}_{-j} when n goes toward infinity. This term is neglected in the following.

Let $X_{\overline{\operatorname{Pa}}(j;\widehat{\mathcal{G}})}$ denote the complementary set of X_j and its parent nodes in $\widehat{\mathcal{G}}$. Then, after Brown et al. (2012), $\frac{1}{n} \sum_{\ell=1}^{n} \log \frac{p(x_{j}^{(\ell)} | \mathbf{X}_{-j}^{(\ell)})}{p(x_{j}^{(\ell)} | x_{n,\ell}^{(\ell)} | \widehat{x}_{0}^{(\ell)})}$ is equal to:

$$\hat{I}^{n}(X_{j}, X_{\overline{\mathbf{Pa}}(j;\widehat{\mathcal{G}})}|X_{\mathbf{Pa}(j;\widehat{\mathcal{G}})}) = \frac{1}{n} \sum_{\ell=1}^{n} \log \frac{p(x_{j}^{(\ell)}, x_{\overline{\mathbf{Pa}}(j;\widehat{\mathcal{G}})}^{(\ell)}|x_{\mathbf{Pa}(j;\widehat{\mathcal{G}})}^{(\ell)})}{p(x_{j}^{(\ell)}|x_{\mathbf{Pa}(j;\widehat{\mathcal{G}})}^{(\ell)})p(x_{\overline{\mathbf{Pa}}(j;\widehat{\mathcal{G}})}^{(\ell)}|x_{\mathbf{Pa}(j;\widehat{\mathcal{G}})}^{(\ell)})},$$
(21)

the estimated conditional mutual information term between X_j and $X_{\overline{Pa}(j;\widehat{\mathcal{C}})}$, conditioned on the

parent variables $X_{\text{Pa}(j;\widehat{\mathcal{G}})}$.

From Eqs (20) and (21) the global loss (Eq. (19)) can be decomposed into a *structural loss* $\mathcal{L}^S(\widehat{\mathcal{G}},D)$ and a *parametric loss* $\mathcal{L}^F(\widehat{\mathcal{G}},\theta^*,D)$:

$$MDL(\widehat{\mathcal{G}}, \theta^*, D) = \mathcal{L}^S(\widehat{\mathcal{G}}, D) + \mathcal{L}^F(\widehat{\mathcal{G}}, \theta^*, D)$$
 (22)

with:

$$\begin{cases} \mathcal{L}^S(\widehat{\mathcal{G}},D) = \sum_{j=1}^d \left[\widehat{I}^n(X_j,X_{\overline{\operatorname{Pa}}(j;\widehat{\mathcal{G}})}|X_{\operatorname{Pa}(j;\widehat{\mathcal{G}})}) \right] + \frac{\lambda_S}{n} |\widehat{\mathcal{G}}| \\ \mathcal{L}^F(\widehat{\mathcal{G}},\theta^*,D) = \sum_{j=1}^d \left[\frac{1}{n} \sum_{\ell=1}^n \log \frac{p(x_j^{(\ell)}|x_{\operatorname{Pa}(j;\widehat{\mathcal{G}})}^{(\ell)})}{q(x_j^{(\ell)}|x_{\operatorname{Pa}(j;\widehat{\mathcal{G}})}^{(\ell)},\theta_j^*)} \right] + \frac{\lambda_F}{n} \|\mathbf{z}_j\|_0 \end{cases}$$

The structural loss $\mathcal{L}^S(\widehat{\mathcal{G}}, D)$, akin category I approaches (Spirtes et al., 2000; Chickering, 2002), aims to identify the Markov equivalence class of the true \mathcal{G} . The parametric loss $\mathcal{L}^F(\widehat{\mathcal{G}}, \theta^*, D)$ instead exploits distribution asymmetries, akin cause effect pair methods (Hoyer et al., 2009; Stegle et al., 2010).

5.1 Identification of the Markov equivalence class with the structural loss

Within the structural loss $\mathcal{L}^S(\widehat{\mathcal{G}},D)$, the minimization of $\hat{I}^n(X_j,X_{\overline{\operatorname{Pa}}(j;\widehat{\mathcal{G}})}|X_{\operatorname{Pa}(j;\widehat{\mathcal{G}})})$ exploits the conditional independence relations in the candidate structure. Let us first consider the case when $\hat{I}^n(X_j, X_{\overline{Pa}(j;\widehat{\mathcal{G}})}|X_{Pa(j;\widehat{\mathcal{G}})})$ is minimized *independently* for each variable X_j (without considering the acyclicity term on $\widehat{\mathcal{G}}$). In the large sample limit and under classical faithfulness and Markov assumptions, Brown et al. (2012) show that the optimum is obtained for $X_{\text{Pa}(i;\widehat{\mathcal{G}})} = MB(X_j)$, the Markov Blanket of X_j in \mathcal{G} . Note that $MB(X_j)$ might contain spurious edges compared to the true parents $X_{Pa(j;\mathcal{G})}$, as it also includes the so-called spouses of X_j : if a child of X_j is retained in $X_{\text{Pa}(i:\widehat{\mathcal{G}})}$, then its parents (spouses) are dependent on X_j conditionally to this child, and are retained

When enforcing the acyclicity of the candidate graph on \widehat{G} and minimizing the structural fitting loss $\mathcal{L}^S(\widehat{\mathcal{G}},D)$

with a regularization term on the total number of edges, spurious edges are removed and the Markov equivalence class of the true DAG (CPDAG) is identified. The intuition is that the acyclicity constraint prevents the children nodes from being selected as parents, hence the spouse nodes do not need be selected either.

In the SAM framework, the CPDAG identification classically relies on the Causal Markov and Faithfulness assumptions (any independence constraint holds in $p(\mathbf{x})$ iff it is present in \mathcal{G}); it also relies on a third assumption on the estimated conditional mutual information bounds.

Theorem 1 (DAG identification up to the Markov equivalence class)

Besides CMA and CFA, let us further assume that for any fixed number of samples n:

- a) for any pair of variables X_i, X_j and any disjoint subset of variables $V \subset \mathbf{X}$, such that $I(X_i, X_i | X_V) = 0$, one has $\hat{I}^n(X_i, X_i | X_V) < \frac{\lambda_S}{n}$.
- b) for any pair of variables X_i, X_j and any disjoint subset of variables $V \subset \mathbf{X}$, such that $I(X_j, X_i | X_V) \neq 0$, one has $\hat{I}^n(X_j, X_i | X_V) > \frac{\lambda_S}{n}$

Then in the limit of large n:

- i) For every $\widehat{\mathcal{G}}$ in the equivalence class of \mathcal{G} , $\mathcal{L}^S(\widehat{\mathcal{G}},D)=\mathcal{L}^S(\mathcal{G},D)$. ii) For every $\widehat{\mathcal{G}}$ not in the equivalence class of \mathcal{G} , $\mathcal{L}^S(\widehat{\mathcal{G}},D)>\mathcal{L}^S(\mathcal{G},D)$.

Proof in Appendix⁷ B

5.2 Identification within Markov equivalence class of DAGs with the parametric loss

The parametric loss $\mathcal{L}^F(\widehat{\mathcal{G}}, \theta^*, D)$ aims to retrieve the true causal model within its Markov equivalence class. Each term

$$\frac{1}{n} \sum_{\ell=1}^{n} \log \frac{p(x_j^{(\ell)} | x_{\operatorname{Pa}(j;\widehat{\mathcal{G}})}^{(\ell)})}{q(x_j^{(\ell)} | x_{\operatorname{Pa}(j;\widehat{\mathcal{G}})}^{(\ell)}, \theta_j^*)}$$

measures the ability of \hat{f}_j to fit the conditional distribution of X_j based on its parents $X_{\text{Pa}}(j;\widehat{\mathcal{G}})$. In the large sample limit, this term converges towards $\mathbb{E}_p\left[\log \frac{p(x_j|x_{\mathrm{Pa}(j;\widehat{\mathcal{G}})})}{q(x_j|x_{\mathrm{Pa}(j;\widehat{\mathcal{G}})},\theta_j^*)}\right]$.

Note that when considering sufficiently powerful causal mechanisms, this term goes to 0 in the large sample limit even if $\mathcal{G} \neq \mathcal{G}$: as shown by Hyvärinen and Pajunen (1999), it is always possible to find a function \hat{f}_j such that $X_j = \hat{f}_j(X_{\text{Pa}(j;\widehat{\mathcal{G}})}, E_j)$, with $E_j \perp \!\!\! \perp X_{\text{Pa}(j;\widehat{\mathcal{G}})}$, corresponding to a probabilistic conditional model q such that $q(x_j|x_{\text{Pa}(j;\widehat{\mathcal{G}})}, \theta_j^*) = p(x_j|x_{\text{Pa}(j;\widehat{\mathcal{G}})})$ (hence

$$\mathbb{E}_p \left[\log \frac{p(x_j | x_{\text{Pa}(j;\widehat{\mathcal{G}})})}{q(x_j | x_{\text{Pa}(j;\widehat{\mathcal{G}})}, \theta_j^*)} \right] = 0).$$

^{7.} The appendix also illustrates this result on the toy 3-variable skeleton A-B-C.

When restricting the *capacity* of the causal mechanism space however, this parametric fitting term may support model identification within the Markov equivalence class of the DAG. Following (Stegle et al., 2010)' pioneering work, SAM uses a soft constraint (a regularization term) to restrict the capacity of the considered mechanism, specifically the number of active neurons involved in \hat{f}_i :

$$\mathcal{L}^F(\widehat{\mathcal{G}}, \theta^*, D) = \frac{1}{n} \sum_{j} \sum_{\ell=1}^{n} \log \frac{p(x_j^{(\ell)} | x_{\text{Pa}(j; \widehat{\mathcal{G}})}^{(\ell)})}{q(x_j^{(\ell)} | x_{\text{Pa}(j; \widehat{\mathcal{G}})}^{(\ell)}, \theta_j^*)} + \lambda_z \|\mathbf{z}_j\|_0$$

Theorem 2 For every DAG $\widehat{\mathcal{G}} \neq \mathcal{G}$ in the Markov equivalence class of \mathcal{G} , given the Working Hypothesis 1 and the causal Markov and faithfulness assumptions:

$$\sum_{j=1}^{d} K(p(x_j|x_{Pa(j;\mathcal{G})})) \stackrel{+}{\leq} \sum_{j=1}^{d} K(p(x_j|x_{Pa(j;\widehat{\mathcal{G}})})), \tag{23}$$

Proof in Appendix C

Following (Janzing and Scholkopf, 2010; Marx and Vreeken, 2017) and approximating the Kolmogorov complexity with the Minimum Description Length (section 3.2), for every DAG $\widehat{\mathcal{G}} \neq \mathcal{G}$ in the Markov equivalence class of \mathcal{G} :

$$MDL(\mathcal{G}, \theta^*, D) \le MDL(\widehat{\mathcal{G}}, \theta^*, D)$$
 (24)

According to equation (22):

$$\mathcal{L}^{S}(\mathcal{G}, D) + \mathcal{L}^{F}(\mathcal{G}, \theta^{*}, D) \leq \mathcal{L}^{S}(\widehat{\mathcal{G}}, D) + \mathcal{L}^{F}(\widehat{\mathcal{G}}, \theta^{*}, D)$$
(25)

Under the conditions given in Theorem 1, for DAGs in the Markov equivalence class of \mathcal{G} in the large sample limit the structural score $\mathcal{L}^S(\widehat{\mathcal{G}}, D)$ is minimal and equal to $\mathcal{L}^S(\mathcal{G}, D)$. It yields:

$$\mathcal{L}^{F}(\mathcal{G}, \theta^*, D) \le \mathcal{L}^{F}(\widehat{\mathcal{G}}, \theta^*, D)$$
(26)

Within the Markov equivalence class, the parametric loss can disambiguate the different structures and support the identification of the true \mathcal{G} . An illustration is presented in Appendix C.

6. Experimental setting

The goal of the validation is to experimentally answer two questions. The first one regards SAM performance compared to the state of the art, depending on whether the underlying joint distribution complies with the usual assumptions (Gaussian distributions for the variables and the noise, linear causal mechanisms). The second question regards the merits and drawbacks of SAM strategy of learning non-linear causal mechanisms, and relying on adversarial learning.

This section first describes the SAM configurations and hyper-parameter settings used in the experiments, followed by the detail of the synthetic, 8 realistic and real-world datasets involved in

^{8.} The codes for generating the synthetic datasets are available at https://github.com/Diviyan-Kalainathan.

the experiments. The baseline algorithms and their hyper-parameter settings, and the performance indicators are last described.

For convenience and reproducibility, all considered algorithms have been integrated in the publicly available CausalDiscovery Toolbox, 9 including the most recent baseline versions at the time of the experiments.

6.1 SAM configurations

Each causal mechanism \hat{f}_j is sought as a 1-hidden layer NN with $n_h^g = 200$ neurons, using tanh activation. Note that this activation function enables to represent linear mechanisms when deemed appropriate. The discriminator is a 2-hidden layer NN with $n_h^D=200$ LeakyReLU units on each layer and batch normalization (Ioffe and Szegedy, 2015). Structural gates $a_{i,j}$ and functional gates $z_{h,j}$ are initialized to 0 with probability 1/2, except for the self-loop terms $a_{i,i}$ set to 0. SAM is trained for $n_{iter} = 10,000$ epochs using Adam (Kingma and Ba, 2014) with initial learning rate 0.01.

SAM hyper-parameters are calibrated using 10 synthetic datasets (five of 20 variables and five of 100 variables) of type VI (section 6.2). In all experiments, $\lambda_S = 5$, $\lambda_F = 0.005$, and

$$\lambda_D = \begin{cases} 0 & \text{if } t < 5,000\\ 1 & \text{otherwise} \end{cases}$$

with t the number of epochs: the first half of the run does not take into account the acyclicity constraint and focuses on the identification of the Markov blankets for each variable; the acyclicity constraint intervenes in the second half of the run.

Four variants have been considered: the full SAM (Alg. 1) and three lesioned variants designed to examine the benefits of non-linear mechanisms and adversarial training.

Specifically, SAM-lin desactivates the non-linear option and only implements linear causal mechanisms (with no functional gates), replacing Eq (7) with:

$$\hat{X}_j = \sum_{i=1}^d W_{j,i} a_{j,i} X_i + W_{j,d+1} E_j + W_{j,0}$$
(27)

A second variant, SAM-mse, replaces the adversarial loss with a standard mean-square error loss, replacing the f-gan term in Eq. (16) with $\frac{1}{n} \sum_{j=1}^{d} \sum_{\ell=1}^{n} (x_{j}^{(\ell)} - \tilde{x}_{j}^{(\ell)})^{2}$. A third variant, **SAM-lin-mse**, involves both linear mechanisms and mean square error losses.

6.2 Benchmarks

The synthetic datasets include 10 DAGs with 20 variables and 10 DAGs with 100 variables.

- 1. The DAG structure is such that the number of parents for each variable is uniformly drawn in $\{0,\ldots,5\};$
- 2. For the *i*-th DAG, the mean μ_i and variance σ_i of the noise variables are drawn as $\mu_i \sim$ $\mathbb{U}(-2,2)$ and $\sigma_i \sim \mathbb{U}(0,0.4)$ and the distribution of the noise variables is set to $\mathcal{N}(\mu_i,\sigma_i)$;

^{9.} https://github.com/diviyan-kalainathan/causaldiscoverytoolbox.

3. For each graph, a 500 sample-dataset is iid generated following the topological order of the graph, with for $\ell=1$ to 500:

$$x^{(\ell)} = (x_1^{(\ell)}, \dots, x_d^{(\ell)}), \quad x_i^{(\ell)} \sim f_i(X_{\text{Pa}(i)}, E_i), \text{ with } E_i \sim \mathcal{N}(\mu_i, \sigma_i)$$

All variables are normalized to zero-mean and unit-variance.

Six categories of causal mechanisms have been considered: besides those considered for the experimental validation of the CAM algorithm (Peters et al., 2014), a more complex one is considered, leveraging the non-linearity of neural nets:

- I. Linear: $X_i = \sum_{j \in Pa(i)} a_{i,j} X_j + E_i$, where $a_{i,j} \sim \mathcal{N}(0,1)$
- II. Sigmoid AM: $X_i = \sum_{j \in \text{Pa}(i)} f_{i,j}(X_j) + E_i$, where $f_{i,j}(x_j) = a \cdot \frac{b \cdot (x_j + c)}{1 + |b \cdot (x_j + c)|}$ with $a \sim \text{Exp}(4) + 1$, $b \sim \mathcal{U}([-2, -0.5] \cup [0.5, 2])$ and $c \sim \mathcal{U}([-2, 2])$.
- III. Sigmoid Mix: $X_i = f_i(\sum_{j \in P_a(j)} X_j + E_i)$, where f_i is as in the previous bullet-point.
- IV. GP AM: $X_i = \sum_{j \in Pa(i)} f_{i,j}(X_j) + E_i$ where $f_{i,j}$ is an univariate Gaussian process with a Gaussian kernel of unit bandwidth.
- V. GP Mix: $X_i = f_i([X_{Pa(i)}, E_i])$, where f_i is a multivariate Gaussian process with a Gaussian kernel of unit bandwidth.
- VI. NN: $X_i = f_i(X_{Pa(i)}, E_i)$, with f_i a 1-hidden layer neural network with 20 tanh units, with all neural weights sampled from $\mathcal{N}(0, 1)$.

6.3 Baseline algorithms

The following algorithms have been used, with their default parameters: the score-based methods GES (Chickering, 2002) and GIES (Hauser and Bühlmann, 2012) with Gaussian scores; the hybrid method MMHC (Tsamardinos et al., 2006), the L_1 penalized method for causal discovery CCDr (Aragam and Zhou, 2015), the LiNGAM algorithm (Shimizu et al., 2006) and the causal additive model CAM (Peters et al., 2014). Lastly, the PC algorithm (Spirtes et al., 2000) has been considered with four conditional independence tests in the Gaussian and non-parametric settings:

- PC-Gauss: using a Gaussian conditional independence test on z-scores;
- PC-HSIC: using the HSIC independence test (Zhang et al., 2012) with a Gamma null distribution (Gretton et al., 2005);
- PC-RCIT: using the Randomized Conditional Independence Test (RCIT) with random Fourier features (Strobl et al., 2017);
- PC-RCOT: the Randomized conditional Correlation Test (RCOT) (Strobl et al., 2017).

PC,¹⁰ GES and LINGAM versions are those of the *pcalg* package (Kalisch et al., 2012). MMHC is implemented with the *bnlearn* package (Scutari, 2009). CCDr is implemented with the *sparsebn* package (Aragam et al., 2017).

^{10.} The better-performing, order-independent version of the PC algorithm proposed by Colombo and Maathuis (2014) is

The GENIE3 algorithm (Irrthum et al., 2010) is also considered, though it does not focus on DAG discovery *per se* as it achieves feature selection, retains the Markov Blanket of each variable using random forest algorithms. Nevertheless, this method won the DREAM4 In Silico Multifactorial challenge (Marbach et al., 2009), and is therefore included in the baseline algorithms (using the *GENIE3* R package).

6.4 Performance indicators

For the sake of robustness, 16 independent runs have been launched for each dataset-algorithm pair. The average causation score $c_{i,j}$ for each edge $X_i \to X_j$ is measured as the fraction of runs where this edge belongs to $\widehat{\mathcal{G}}$. When an edge is left undirected, e.g with PC algorithm, it is counted as appearing with both orientations with weight 1/2.

Precision-recall A true positive is an edge $i \to j$ of the true DAG $\mathcal G$ which is correctly recovered by the algorithm; T_p is the number of true positive. A false negative is an edge of $\mathcal G$ which is missing in $\widehat{\mathcal G}$; F_n is the number of false negatives. A false positive is an edge in $\widehat{\mathcal G}$ which is not in $\mathcal G$ (reversed edges and edges which are not in the skeleton of $\mathcal G$); F_p is the number of false positives. The precision-recall curve, showing the tradeoff between precision $(T_p/(T_p+F_p))$ and recall $(T_p/(T_p+F_n))$ for different causation thresholds (Fig. 7), is summarized by the **Area under the Precision Recall Curve (AuPR)**, ranging in [0,1], with 1 being the best. 11

Structural Hamming Distance Another performance indicator used in the causal graph discovery framework is the Structural Hamming Distance (SHD) (Tsamardinos et al., 2006), set to the number of missing edges and redundant edges in the found structure. This SHD score is computed in the following by considering all edges $i \to j$ with $c_{i,j} > .5$. Note that a reversal error (retaining $j \to i$ while \mathcal{G} includes edge $i \to j$) is counted as a single mistake.

$$SHD(\hat{A}, A) = \sum_{i,j} |\hat{A}_{i,j} - A_{i,j}| - \frac{1}{2} \sum_{i,j} (1 - \max(1, \hat{A}_{i,j} + A_{j,i})), \tag{28}$$

with A (respectively \hat{A}) the adjacency matrix of \mathcal{G} (resp. the found causal graph $\hat{\mathcal{G}}$).

7. Experiments

This section first reports on the experimental results obtained on synthetic datasets. Realistic biological data coming from the SynTREN simulator (Van den Bulcke et al., 2006) on 20- and 100-node graphs, and from GeneNetWeaver (Schaffter et al., 2011) on the DREAM4 challenge are thereafter considered (section 7.2), and we last consider the extensively studied flow cytometry dataset (Sachs et al., 2005) (section 7.3). A t-test is used to assess whether the score difference between any two methods is statistically significant with a p-value below 0.001.

The detail of all results is given in Appendix D, reporting the average performance indicators, standard deviation, and computational cost of all considered algorithms.

7.1 Synthetic datasets

11. Using the *scikit-learn v0.20.1* library (Pedregosa et al., 2011).

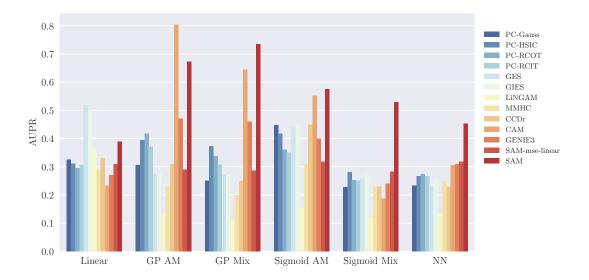


Figure 4: Performance of causal graph discovery methods on 20-node synthetic graphs measured by the Area under the Precision Recall Curve (the higher, the better). SAM ranks among the top-three methods, being only dominated by GES and GIES for linear mechanisms and by CAM for univariate mechanisms (better seen in color).

20 variable-graphs The comparative results (Fig. 4) demonstrate SAM robustness in term of Area under the Precision Recall Curve (AUPR) on all categories of 20-node graphs. Specifically, SAM is dominated by GES and GIES on linear mechanisms and by CAM for Gaussian univariate mechanisms, reminding that GES and GIES (resp. CAM) specifically aim at linear mechanisms (resp. Gaussian univariate mechanisms). Note that, while the whole ranking of the algorithms may depend on the considered performance indicator, the best performing algorithm is most often the same regardless of whether the AUPR or the Structural Hamming distance is considered. For non-linear cases with complex interactions (the Sigmoid Mix and NN cases), SAM significantly outperforms other non-parametric methods such as PC-HSIC, PC-RCOT and PC-RCIT. In the linear Gaussian setting, SAM aims to the Markov equivalence class of the true graph (under causal Markov and faithfulness assumptions) and performs less well than for e.g. the GP mix where SAM can exploit both conditional independence relations and distribution asymmetries. Though seemingly counter-intuitive, a graph with more complex interactions between noise and variables may be actually easier to recover than a graph generated with simple mechanisms (see also Wang and Blei (2018)).

SAM computational cost is one order of magnitude higher than that of the other methods (all measured on a single CPU core Intel Xeon 2.7Ghz).¹² The lesioned versions, SAM-lin, SAM-mse and SAM-line-mse have significantly worse performances than SAM (except for the linear

^{12.} A speed up factor of 25 can be obtained for SAM using a GPU environment with single graphic card GeForce GTX 1080Ti, particularly beneficial for the GAN training.

mechanism and additive Gaussian noise cases), demonstrating the merits of the NN-based and adversarial learning approach in the general case.

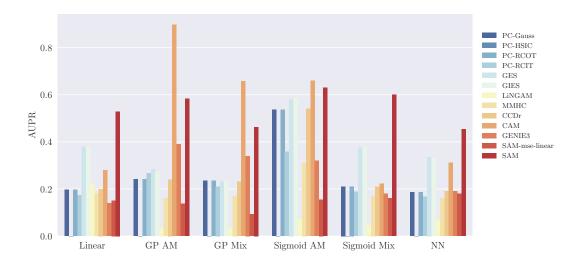


Figure 5: Performance of causal graph discovery methods on 100-node synthetic graphs measured by the Area under the Precision Recall Curve (the higher, the better). On datasets relying on Gaussian processes, CAM tops the leaderboard by a significant margin as its search space matches the sought causal mechanisms. SAM demonstrates its robustness with respect to the underlying generative models (better seen in color).

100-variable graphs The comparative results on the 100-node graphs (Fig. 5) confirm the good overall robustness of SAM. As could have been expected, SAM is dominated by CAM on the GP AM, GP Mix and Sigmoid AM; indeed, focusing on the proper causal mechanism space yields a significant advantage, all the more so as the number of variables increases. Nevertheless, SAM does never face a catastrophic failure, and it even performs quite well on linear datasets. A tentative explanation is based on the fact that the *tanh* activation function enables to capture linear mechanisms; another explanation is based on the adversarial loss, empirically more robust than the MSE loss in high-dimensional problems.

In terms of computational cost, SAM scales well at d=100 variables, particularly when compared to its best competitor CAM, that uses a combinatorial graph search. The PC-HSIC algorithm had to be stopped after 50 hours; more generally, constraint-based methods based on the PC algorithm do not scale well w.r.t. the number of variables.

7.2 Simulated biological datasets

As said, the SynTREN (Van den Bulcke et al., 2006) and GeneNetWeaver (GNW) (Schaffter et al., 2011) simulators of genetic regulatory networks have been used to generate observational data reflecting realistic complex regulatory mechanisms, high-order conditional dependencies between expression patterns and potential feedback cycles, based on an available causal model.

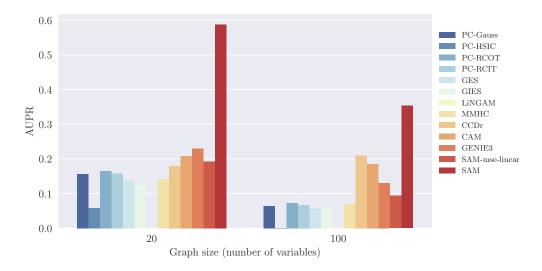


Figure 6: Performance of causal graph discovery methods on SynTREN graphs measured by the Area under the Precision Recall Curve (the higher, the better). Left: 20 nodes. Right: 100 nodes (better seen in color).

SynTREN simulator Sub-networks of E. coli (Shen-Orr et al., 2002) have been considered, where interaction kinetics are based on Michaelis-Menten and Hill kinetics (Mendes et al., 2003). Overall, ten 10-nodes and ten 100-nodes graphs have been considered. For each graph, 500-sample datasets are generated by SynTREN.

Likewise, the comparative results on all SynTREN graphs (Fig. 6) demonstrate the good performances of SAM. Overall, the best performing methods take into account both distribution asymmetry and multivariate interactions. Constraint-based methods are hampered by the lack of v-structures, preventing the orientation of many edges to be based on CI tests only (PC-HSIC algorithm was stopped after 50 hours and LiNGAM did not converge on one of the datasets). The benefits of using non-linear mechanisms on such problems are evidenced by the difference between SAM-lin-mse and SAM-mse (Appendix D). The Precision-Recall curve is displayed on Fig. 7 for representative 20-node and 100-node graphs, confirming that SAM can be used to infer networks having complex distributions, complex causal mechanisms and interactions.

GeneNetWeaver simulator - DREAM4 Five 100-nodes graphs generated using the GeneNetWeaver simulator define the *In Silico Size 100 Multifactorial* challenge track of the *Dialogue for Reverse Engineering Assessments and Methods* (DREAM) initiative. These graphs are subnetworks of transcriptional regulatory networks of E. coli and S. cerevisiae and their dynamics are simulated using a kinetic gene regulation model, where noise is added both in the dynamics of the networks and on the measurement of expression data. Multifactorial perturbations are simulated by slightly increasing

^{13.} Random seeds set to 1...10 are used for the sake of reproducibility. SynTREN hyper-parameters include a probability of 1.0 (resp. 0.1) for complex 2-regulator interactions (resp. for biological noise, experimental noise and noise on correlated inputs).

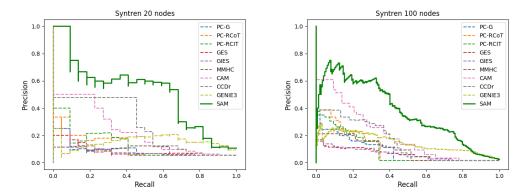


Figure 7: Precision/Recall curve for two SynTREN graphs: Left, 20 nodes; Right, 100 nodes (better seen in color).

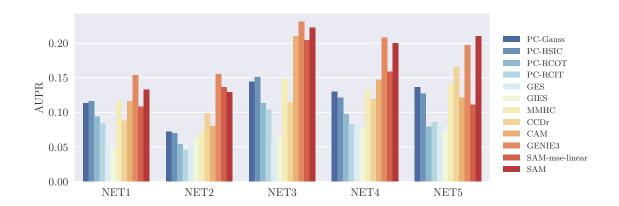


Figure 8: Performance of causal graph discovery methods on 5 artificial datasets of the Dream4 In Silico Multifactorial Challenge measured by the Area under the Precision Recall Curve (the higher, the better). GENIE3 achieves the best performance on 4 datasets, with SAM close second (better seen in color).

or decreasing the basal activation of all genes of the network simultaneously by different random amounts. In total, the number of expression conditions for each network is set to 100.

The comparative results on these five graphs (Fig. 8) show that GENIE3 outperforms all other methods, with SAM ranking second. A tentative explanation for GENIE3 excellent performance is that it does not enforce the discovery of acyclic graphs, which is appropriate as regulatory networks involve feedback loops. The Precision/Recall curves (Fig. 9) demonstrate that SAM matches GENIE3 performances in the low recall region. Overall, on such complex problem domains, it appears relevant to make few assumptions on the underlying generative model (like GENIE3 and SAM), while being able to capture high-order conditional dependencies between variables. Note that LiNGAM did not converge on one of these datasets.

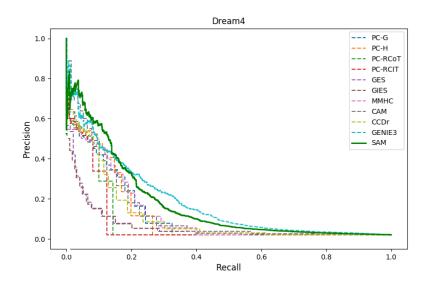


Figure 9: Precision/Recall curve for the Dream4 *In Silico Multifactorial Challenge* (better seen in color).

7.3 Real-world biological data

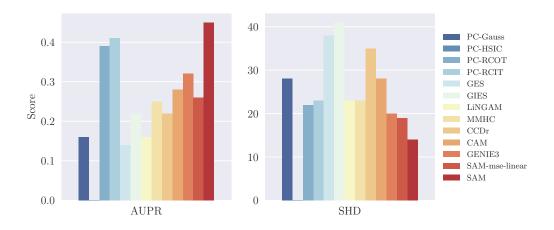


Figure 10: Performance of causal graph discovery methods on the protein network problem (Sachs et al., 2005). Left, Area under the Precision Recall curve (the higher the better); Right, Structural Hamming distance (the lower, the better). SAM significantly outperforms all other methods on this dataset (better seen in color).

The well-studied protein network problem (Sachs et al., 2005) is associated with observational data including 7,466 observational samples. Same experimental setting is used as for the other problem, with a bootstrap ratio of 0.8. According to both performance indicators (Fig. 10), SAM

significantly outperforms the other methods. The precision/recall curve (Fig. 11) shows that SAM is particularly accurate when its confidence score is high, showing that for critical applications where false negatives are to be avoided, using SAM with a threshold is a viable option.

Notably, SAM recovers the transduction pathway $raf \rightarrow mek \rightarrow erk$ corresponding to direct enzyme-substrate causal effect (Sachs et al., 2005).

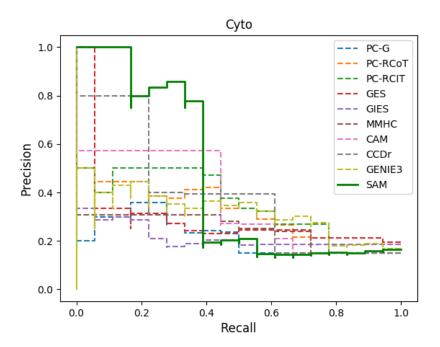


Figure 11: Precision/Recall curve for the curve protein network (better seen in color).

8. Discussion and Perspectives

The main contribution of the paper is a unifying causal discovery framework, exploiting both structural independence and distributional asymmetries through optimizing well-founded structural and functional criteria. This framework is implemented in the SAM algorithm¹⁴, leveraging the non-parametric power of Generative Adversarial Neural networks (GANs) to capture a faithful generative model and enforce the discovery of acyclic causal graphs through sparsity and algebraic regularizations, using stochastic gradient descent.

Extensive empirical evidence is gathered to show SAM robustness across diverse synthetic, realistic and real-world problems. Lesion studies are conducted to assess whether and when it is beneficial to learn non-linear mechanisms and to rely on adversarial learning as opposed to MSE minimization.

As could have been expected, in particular settings SAM is dominated by algorithms specifically designed for this setting, such as CAM (Bühlmann et al., 2014) in the case of additive noise model and Gaussian process mechanisms, and GENIE3 when facing causal graphs with feedback loops.

^{14.} Available at https://github.com/Diviyan-Kalainathan/SAM.

Nevertheless, SAM most often ranks first and always avoids catastrophic failures. The main limitation of SAM is its computational cost, higher by an order of magnitude than other approaches on 20-variable problems. On 100-variable problems however, SAM catches up with the other approaches as it avoids the combinatorial exploration of the graph space.

This work opens up four avenues for further research. An on-going extension regards the case of categorical and mixed variables, taking inspiration from discrete GANs (Hjelm et al., 2017). Another perspective is to relax the causal sufficiency assumption and handle hidden confounders, e.g. by introducing statistical dependencies between the noise variables attached to different variables (Rothenhäusler et al., 2015), or creating shared noise variables (Janzing and Schölkopf, 2018), or via dimensionality reduction (Wang and Blei, 2018). A longer term perspective is to extend SAM to simulate interventions on target variables. Lastly, the case of causal graphs with cycles will be considered, leveraging the power of recurrent neural nets to define a proper generative model from a graph with feedback loops.

References

- Bryon Aragam and Qing Zhou. Concave penalized estimation of sparse gaussian bayesian networks. *Journal of Machine Learning Research*, 16:2273–2328, 2015.
- Bryon Aragam, Jiaying Gu, and Qing Zhou. Learning large-scale bayesian networks with the sparsebn package. *arXiv preprint arXiv:1703.04025*, 2017.
- Andrew R Barron and Thomas M Cover. Minimum complexity density estimation. *IEEE transactions on information theory*, 37(4):1034–1054, 1991.
- David A Bell and Hui Wang. A formalism for relevance and its application in feature subset selection. *Machine learning*, 41(2):175–195, 2000.
- Gavin Brown, Adam Pocock, Ming-Jie Zhao, and Mikel Luján. Conditional likelihood maximisation: a unifying framework for information theoretic feature selection. *Journal of machine learning research*, 13(Jan):27–66, 2012.
- Kailash Budhathoki and Jilles Vreeken. Causal inference by stochastic complexity. *arXiv preprint arXiv:1702.06776*, 2017.
- Peter Bühlmann, Jonas Peters, Jan Ernest, et al. Cam: Causal additive models, high-dimensional order search and penalized regression. *The Annals of Statistics*, 42(6):2526–2556, 2014.
- Pu Chen, Chihying Hsiao, Peter Flaschel, and Willi Semmler. Causal analysis in economics: Methods and applications. 2007.
- David Maxwell Chickering. Optimal structure identification with greedy search. *JMLR*, 2002.
- Diego Colombo and Marloes H Maathuis. Order-independent constraint-based causal structure learning. *JMLR*, 2014.
- F. Doshi-Velez and B. Kim. Towards a rigorous science of interpretable machine learning. *arXiv:1702.08608*, 2017.

- Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *NIPS*, 2014.
- Olivier Goudet, Diviyan Kalainathan, Philippe Caillou, Isabelle Guyon, David Lopez-Paz, and Michele Sebag. Learning functional causal models with generative neural networks. In *Explainable and Interpretable Models in Computer Vision and Machine Learning*, pages 39–80. Springer, 2018.
- Arthur Gretton, Ralf Herbrich, Alexander Smola, Olivier Bousquet, and Bernhard Schölkopf. Kernel methods for measuring independence. *JMLR*, 2005.
- Peter D Grünwald. The minimum description length principle. MIT press, 2007.
- Peter D Grünwald, Paul MB Vitányi, et al. Algorithmic information theory. *Handbook of the Philosophy of Information*, pages 281–320, 2008.
- Alain Hauser and Peter Bühlmann. Characterization and greedy learning of interventional markov equivalence classes of directed acyclic graphs. *Journal of Machine Learning Research*, 13(Aug): 2409–2464, 2012.
- R Devon Hjelm, Athul Paul Jacob, Tong Che, Adam Trischler, Kyunghyun Cho, and Yoshua Bengio. Boundary-seeking generative adversarial networks. *arXiv preprint arXiv:1702.08431*, 2017.
- Patrik O Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. *NIPS*, 2009.
- Aapo Hyvärinen and Petteri Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural Networks*, 12(3):429–439, 1999.
- Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.
- Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456, 2015.
- Alexandre Irrthum, Louis Wehenkel, Pierre Geurts, et al. Inferring regulatory networks from expression data using tree-based methods. *PloS one*, 5(9):e12776, 2010.
- Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv* preprint arXiv:1611.01144, 2016.
- Dominik Janzing and Bernhard Scholkopf. Causal inference using the algorithmic markov condition. *IEEE Transactions on Information Theory*, 56(10):5168–5194, 2010.
- Dominik Janzing and Bernhard Schölkopf. Detecting confounding in multivariate linear models via spectral analysis. *Journal of Causal Inference*, 6(1), 2018.
- Markus Kalisch, Martin Mächler, Diego Colombo, Marloes H Maathuis, Peter Bühlmann, et al. Causal inference using graphical models with the r package pealg. *Journal of Statistical Software*, 2012.

- Durk P Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. ICLR, 2014.
- Jan Lemeire and Kris Steenhaut. Inference of graphical causal models: Representing the meaningful information of probability distributions. In *Causality: Objectives and Assessment*, pages 107–120, 2010.
- Philippe Leray and Patrick Gallinari. Feature selection with neural networks. *Behaviormetrika*, 26 (1):145–166, 1999.
- Ming Li and Paul Vitányi. *An introduction to Kolmogorov complexity and its applications*. Springer Science & Business Media, 2013.
- David Lopez-Paz, Krikamol Muandet, Bernhard Schölkopf, and Ilya O Tolstikhin. Towards a learning theory of cause-effect inference. *ICML*, 2015.
- Christos Louizos, Max Welling, and Diederik P Kingma. Learning sparse neural networks through *l*₋0 regularization. *arXiv preprint arXiv:1712.01312*, 2017.
- Chris J Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv* preprint arXiv:1611.00712, 2016.
- Daniel Marbach, Thomas Schaffter, Dario Floreano, Robert J Prill, and Gustavo Stolovitzky. The dream4 in-silico network challenge. *Draft, version 0.3*, 2009.
- Alexander Marx and Jilles Vreeken. Telling cause from effect using mdl-based local and global regression. In 2017 IEEE international conference on data mining (ICDM), pages 307–316. IEEE, 2017.
- Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. *The annals of statistics*, 2006.
- Pedro Mendes, Wei Sha, and Keying Ye. Artificial gene networks for objective comparison of analysis algorithms. *Bioinformatics*, 19(suppl_2):ii122–ii129, 2003.
- Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. arXiv, 2014.
- Joris M Mooij, Jonas Peters, Dominik Janzing, Jakob Zscheischler, and Bernhard Schölkopf. Distinguishing cause from effect using observational data: methods and benchmarks. *JMLR*, 2016.
- XuanLong Nguyen, Martin J Wainwright, and Michael I Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, 2010.
- Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. In *Advances in Neural Information Processing Systems*, pages 271–279, 2016.
- Judea Pearl. Causality: models, reasoning, and inference. *Econometric Theory*, 19(675-685):46, 2003a.
- Judea Pearl. Causality: models, reasoning and inference. *Econometric Theory*, 2003b.

- Judea Pearl. Causality. 2009.
- Judea Pearl and Thomas Verma. A formal theory of inductive causation. 1991.
- F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
- Jonas Peters, Joris M Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. *The Journal of Machine Learning Research*, 15(1):2009–2053, 2014.
- Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference Foundations and Learning Algorithms*. MIT Press, 2017.
- John A Quinn, Joris M Mooij, Tom Heskes, and Michael Biehl. Learning of causal relations. *ESANN*, 2011.
- Zhiquan Ren, Yang Yang, Feng Bao, Yue Deng, and Qionghai Dai. Directed adaptive graphical lasso for causality inference. *Neurocomputing*, 173:1989–1994, 2016.
- Dominik Rothenhäusler, Christina Heinze, Jonas Peters, and Nicolai Meinshausen. Backshift: Learning causal cyclic graphs from unknown shift interventions. In *Advances in Neural Information Processing Systems*, pages 1513–1521, 2015.
- Karen Sachs, Omar Perez, Dana Pe'er, Douglas A Lauffenburger, and Garry P Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721): 523–529, 2005.
- Thomas Schaffter, Daniel Marbach, and Dario Floreano. Genenetweaver: in silico benchmark generation and performance profiling of network inference methods. *Bioinformatics*, 27(16): 2263–2270, 2011.
- Marco Scutari. Learning bayesian networks with the bnlearn r package. arXiv, 2009.
- Shai S Shen-Orr, Ron Milo, Shmoolik Mangan, and Uri Alon. Network motifs in the transcriptional regulation network of escherichia coli. *Nature genetics*, 31(1):64, 2002.
- Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, and Antti Kerminen. A linear non-gaussian acyclic model for causal discovery. *JMLR*, 2006.
- Ali Shojaie and George Michailidis. Penalized likelihood methods for estimation of sparse high-dimensional directed acyclic graphs. *Biometrika*, 97(3):519–538, 2010.
- Peter Spirtes and Kun Zhang. Causal discovery and inference: concepts and recent methodological advances. *Applied informatics*, 2016.
- Peter Spirtes, Clark N Glymour, and Richard Scheines. Causation, prediction, and search. 2000.

- Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.
- Oliver Stegle, Dominik Janzing, Kun Zhang, Joris M Mooij, and Bernhard Schölkopf. Probabilistic latent variable models for distinguishing between cause and effect. *NIPS*, 2010.
- Eric V Strobl, Kun Zhang, and Shyam Visweswaran. Approximate kernel-based conditional independence tests for fast non-parametric causal discovery. *arXiv* preprint arXiv:1702.03877, 2017.
- Ioannis Tsamardinos, Laura E Brown, and Constantin F Aliferis. The max-min hill-climbing bayesian network structure learning algorithm. *Machine learning*, 65(1):31–78, 2006.
- Tim Van den Bulcke, Koenraad Van Leemput, Bart Naudts, Piet van Remortel, Hongwu Ma, Alain Verschoren, Bart De Moor, and Kathleen Marchal. Syntren: a generator of synthetic gene expression data for design and analysis of structure learning algorithms. *BMC bioinformatics*, 7 (1):43, 2006.
- Jorge R Vergara and Pablo A Estévez. A review of feature selection methods based on mutual information. *Neural computing and applications*, 24(1):175–186, 2014.
- Yixin Wang and David M. Blei. The blessings of multiple causes. *CoRR*, abs/1805.06826, 2018. URL http://arxiv.org/abs/1805.06826.
- Kui Yu, Lin Liu, and Jiuyong Li. A unified view of causal and non-causal feature selection. *arXiv* preprint arXiv:1802.05844, 2018.
- Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. *arXiv*, 2012.
- Xun Zheng, Bryon Aragam, Pradeep Ravikumar, and Eric P. Xing. Dags with NO TEARS: continuous optimization for structure learning. In *NeurIPS*, pages 9492–9503, 2018a.
- Xun Zheng, Bryon Aragam, Pradeep Ravikumar, and Eric P Xing. Dags with no tears: Smooth optimization for structure learning. *arXiv preprint arXiv:1803.01422*, 2018b.

Appendices

A. Notations and definitions

Notation	Definition
X	Set of continuous random variables X_1, \ldots, X_d
$\mathbf{X}_{ackslash i,j}$	Set of all continuous random variables in X except X_i and X_j
D	iid n -sample of X
x_j^l	l -th sample of X_i
$p(x_j)$	True marginal probability density function of X_j
$p(x_j x_i)$	<i>True</i> conditional probability density function of X_i conditionally to X_i
$p(\mathbf{x})$	<i>True</i> joint probability density function of X
$q(x_j)$	Generated marginal probability density function for X_j
$q(x_j x_i)$	Generated conditional probability density function of X_j conditionally to X_i
$q(\mathbf{x})$	Generated joint probability density function for X
\mathcal{G}	True causal graph associated to \mathbf{X} ; X_j is both a continuous random variable and a node
	in ${\mathcal G}$
$\hat{\mathcal{G}}$	Candidate causal graph
$ \mathcal{G} $	Total number of edges in \mathcal{G}
$X_{\operatorname{Pa}(j;\mathcal{G})}$	Set of parents of the X_j node in \mathcal{G}
$X_{\overline{\operatorname{Pa}}(j;\mathcal{G})}$	Set of variables that are not parents of X_j in \mathcal{G} nor X_j itself
$H(X_i)$	Entropy of variable X_i
$I(X_i, X_j)$	Mutual Information between X_i and X_j
$I(X_i, X_j X_k)$	Conditional mutual Information between X_i and X_j conditionally to X_k
$ ho_{i,j}$	Pearson correlation coefficient between X_i and X_j
$D_{KL}(p(\mathbf{x}) \parallel q(\mathbf{x}))$	Kullback-Leibler Divergence between the joint probability density functions p and q of
(2 () 2 () /	X
θ	Set of parameters of a SAM (except the functional and structural gates z_{ij} , a_{ij})
$ heta^*$	Optimal set of parameters θ of a SAM, that minimises the loss in a given configuration
$X_i \perp \!\!\! \perp X_j \mathbf{X}_{i,j}$	Variables X_i and X_j are independent conditionally to all other variables in X
$\mathrm{MB}(X_i)$	Markov blanket of the variable (node) X_i
Σ	Covariance matrix of X
S	Covariance matrix of D
K	Precision matrix of X
FCM	Functional Causal Model
DAG	Directed Acyclic Graph
CPDAG	Completed Partially Directed Acyclic Graph
CMA	Causal Markov Assumption
CFA	Causal Faithfulness Assumption
CSA	Causal Sufficiency Assumption

Table 1: Notations used throughout the paper

B. Structural loss: Proof of Theorem 1 and example

Theorem 1 [DAG identification up to the Markov equivalence class]

Besides CMA and CFA assumptions, it is further assumed that for any fixed number of samples n: a) for any pair of variables X_i, X_j and any disjoint set of variables $V \subset \mathbf{X}$, such that $I(X_j, X_i | X_V) = 0$, its empirical counterpart estimated with the data sample, $\hat{I}^n(X_j, X_i | X_V)$, is below $\frac{\lambda_S}{n}$. b) for any pair of variables X_i, X_j and any disjoint set of variables $V \subset \mathbf{X}$, such that $I(X_j, X_i | X_V) \neq 0$, $\hat{I}^n(X_j, X_i | X_V)$ is above $\frac{\lambda_S}{n}$.

Then in the limit of large n:

- i) For every $\widehat{\mathcal{G}}$ in the equivalence class of \mathcal{G} , $\mathcal{L}^S(\widehat{\mathcal{G}},D)=\mathcal{L}^S(\mathcal{G},D)$. ii) For every $\widehat{\mathcal{G}}$ not in the equivalence class of \mathcal{G} , $\mathcal{L}^S(\widehat{\mathcal{G}},D)>\mathcal{L}^S(\mathcal{G},D)$.

Proof Let $\widehat{\mathcal{G}}$ be a DAG, and let $\widehat{\mathcal{G}}'$ be defined from $\widehat{\mathcal{G}}$ by adding a single edge $X_k \to X_j$ such that $\widehat{\mathcal{G}}'$ is still a DAG. Let us compare the structural losses of $\widehat{\mathcal{G}}$ and $\widehat{\widehat{\mathcal{G}}}'$:

$$\begin{split} \Delta \mathcal{L}^S &= \mathcal{L}^S(\widehat{\mathcal{G}}', D) - \mathcal{L}^S(\widehat{\mathcal{G}}, D) \\ &= \widehat{I}^n(X_{\overline{\operatorname{Pa}}(j; \widehat{\mathcal{G}}')}, X_j | X_{\operatorname{Pa}(j; \widehat{\mathcal{G}}')}) - \widehat{I}^n(X_{\overline{\operatorname{Pa}}(j; \widehat{\mathcal{G}})}, X_j | X_{\operatorname{Pa}(j; \widehat{\mathcal{G}})}) + \frac{\lambda_S}{n} \end{split}$$

From

$$\hat{I}^{n}(X_{j}, X_{-j}) = \hat{I}^{n}(X_{j}, X_{\text{Pa}(j;\widehat{\mathcal{G}}')}) + \hat{I}^{n}(X_{\overline{\text{Pa}}(j;\widehat{\mathcal{G}}')}, X_{j} | X_{\text{Pa}(j;\widehat{\mathcal{G}}')})$$
(29)

and

$$\hat{I}^n(X_j, X_{-j}) = \hat{I}^n(X_j, X_{\operatorname{Pa}(j:\widehat{G})}) + \hat{I}^n(X_{\overline{\operatorname{Pa}}(j:\widehat{G})}, X_j | X_{\operatorname{Pa}(j:\widehat{G})})$$
(30)

it follows:

$$\Delta \mathcal{L}_{S} = -\hat{I}^{n}(X_{j}, X_{\operatorname{Pa}(j;\widehat{\mathcal{G}}')}) + \hat{I}^{n}(X_{j}, X_{\operatorname{Pa}(j;\widehat{\mathcal{G}})}) + \frac{\lambda_{S}}{n}$$

$$= -\hat{I}^{n}(X_{j}, X_{\operatorname{Pa}(j;\widehat{\mathcal{G}})} \cup X_{k}) + \hat{I}^{n}(X_{j}, X_{\operatorname{Pa}(j;\widehat{\mathcal{G}})}) + \frac{\lambda_{S}}{n}$$

$$= -\hat{I}^{n}(X_{j}, X_{k} | X_{\operatorname{Pa}(j;\widehat{\mathcal{G}})}) + \frac{\lambda_{S}}{n}$$

- $\bullet \ \ \text{If } X_j \bot \!\!\! \bot X_k | X_{\operatorname{Pa}(j;\widehat{\mathcal{G}})}, \text{ then } I(X_j, X_k | X_{\operatorname{Pa}(j;\widehat{\mathcal{G}})}) = 0 \text{ and according to clause a), } \\ \hat{I}^n(X_j, X_k | X_{\operatorname{Pa}(j;\widehat{\mathcal{G}})}) < 0 \text{ and } C_j \cap C_j \cap$ $\frac{\lambda_S}{n}$ and $\Delta \mathcal{L}^S > 0$. In other words, in the sample size limit the loss increases when adding any irrelevant edge.
- If $X_j \not\perp \!\!\! \perp X_k | X_{\operatorname{Pa}(j;\widehat{\mathcal{G}})}$, then $I(X_j, X_k | X_{\operatorname{Pa}(j;\widehat{\mathcal{G}})}) \neq 0$. It follows from the clause b) that $\hat{I}^n(X_j, X_k | X_{\text{Pa}(j;\widehat{\mathcal{G}})}) > \frac{\lambda_S}{n}$ and therefore $\Delta \mathcal{L}^S < 0$. Likewise, the loss decreases when adding any edge that removes an irrelevant conditional independence.

Both results establish the consistency of the structural loss \mathcal{L}^S : the DAGs minimizing the structural loss belong to the Markov equivalence class of \mathcal{G} (Chickering (2002), Prop 8).

Illustration with toy data that SAM identifies V-structures based on the structural loss. We illustrate on a toy example that SAM can orient causal arrows based solely on Markov properties of the data distributions. To that end we use a simple example of V-structures in which functional dependencies are linear and the noise Gaussian. The choice of linear dependency and Gaussian noise makes it impossible to use distributional asymmetries other than conditional independence (Shimizu et al., 2006; Hoyer et al., 2009).

We consider a triplet of variables (A, B, C). With no loss of generality, the graph skeleton involving variables (A, B, C) is A - B - C. All three causal models (up to variable renaming) based on this skeleton are used to generate 1000-sample datasets, where the random noise variables are independent centered Gaussian variables.

Given skeleton A - B - C, the four possible graph structures are fitted with SAM based on data generated with the V-structure graph:

- \mathcal{L}_{ABC} and \mathcal{L}_{CBA} : Chain structures $A \to B \to C$ and $A \leftarrow B \leftarrow C$,
- $\mathcal{L}_{Vstruct}$ (V structure): $A \to B \leftarrow C$
- \mathcal{L}_{revV} (reversed V structure): $A \leftarrow B \rightarrow C$

where \mathcal{L}_{ABC} , \mathcal{L}_{CBA} , $\mathcal{L}_{Vstruct}$ and \mathcal{L}_{revV} denote the resulting loss of the SAM models respectively attached to these structures after training, obtained by setting the graph constant and the mechanisms linear. We used the experimental setting of SAM detailed in Section 6.1, with the Linear variant SAM-lin in order to avoid the influences from the mechanism regularization.

The fit losses measured by the discriminator (cf. Section 4.2) and computed on all three datasets are displayed in Figure 12 (average over 128 runs).

SAM's loss supports a clear and significant discrimination between the V-structure and all other structures thus showing that SAM can effectively detect, and leverage conditional independencies between variables.

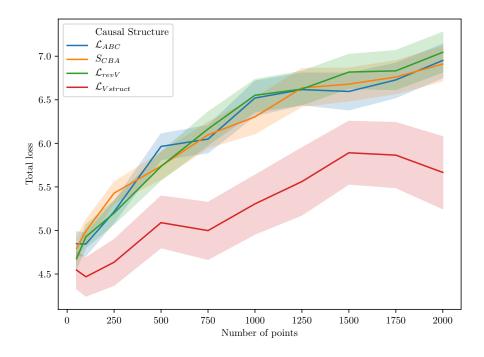


Figure 12: **Identification of V-structures:** In these experiments, the causal direction is identifiable only from the graph structure (not from distributional asymmetries, since we use linear dependencies and Gaussian noise). We use a dataset of V-structure graphs. We show that V-structure graphs have always a better (lower) fitness loss than other structures independently from the number of points in the dataset.

C. Parametric loss: Proof of Theorem 2 and example

Theorem 2

For every DAG $\widehat{\mathcal{G}} \neq \mathcal{G}$ in the Markov equivalence class of \mathcal{G} , given the Working Hypthesis 1 and the causal Markov and faithfulness assumption:

$$\sum_{j=1}^{d} K(p(x_j|x_{Pa(j;\mathcal{G})})) \stackrel{+}{\leq} \sum_{j=1}^{d} K(p(x_j|x_{Pa(j;\widehat{\mathcal{G}})})), \tag{31}$$

Proof The proof is given by observing that the Working Hypothesis 1 implies that the shortest description of p is given by separate descriptions of the conditional probability distributions:

$$K(p(\mathbf{x})) \stackrel{+}{=} \sum_{j=1}^{d} K(p(x_j|x_{\text{Pa}(j;\mathcal{G})})), \tag{32}$$

This equality holds if the conditionals $p(x_j|x_{Pa(j;\mathcal{G})})$ are assumed to be algorithmically independent (Janzing and Scholkopf, 2010).

For any DAGs $\widehat{\mathcal{G}} \neq \mathcal{G}$ in the Markov equivalence class of the \mathcal{G} :

$$p(\mathbf{x}) = \prod_{j=1}^{d} p(x_j | x_{\text{Pa}(j;\widehat{\mathcal{G}})})$$
(33)

According to Lemeire and Steenhaut (2010), the sum of the description of the conditionals $p(x_j|x_{\text{Pa}(j;\widehat{\mathcal{G}})})$ is always greater than the description of their product. It gives for every DAG $\widehat{\mathcal{G}} \neq \mathcal{G}$ in the Markov equivalence class of \mathcal{G} :

$$K(p(\mathbf{x})) \stackrel{+}{\leq} \sum_{j=1}^{d} K(p(x_j | x_{\operatorname{Pa}(j;\widehat{\mathcal{G}})}))$$
(34)

Thus:

$$\sum_{j=1}^{d} K(p(x_j|x_{\text{Pa}(j;\mathcal{G})})) \stackrel{+}{\leq} \sum_{j=1}^{d} K(p(x_j|x_{\text{Pa}(j;\widehat{\mathcal{G}})})), \tag{35}$$

Illustration with toy data of Markov equivalence class disambiguation. The SAM architecture leverages model complexity to distinguish between cause and effect. To validate this point, we apply SAM on the following dataset with only two variables X and Y (a known case of a Markov equivalence class X-Y):

$$\begin{cases} X \sim U(-1,1) \\ E_y \sim U(-.33,.33) \\ Y = 4(X^2 - 0.5)^2 + E_y \end{cases}$$

The figure 13 represents this two-dimensional distribution. Interestingly enough the Pearson

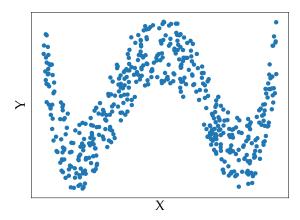


Figure 13: Scatter plot of 500 points sampled from the distribution (X, Y)

correlation coefficient between X and Y is equal to 0, showing the need for a non-linear statistical model in order to explain the relation between the two variables. Methods leveraging conditional independencies are unable to identify the causal direction in this case, as only two variables are available. In this case, the FCM corresponding to the causal direction $X \to Y$ seems simpler than the anticausal direction $Y \to X$. Moreover the residual of the non-linear regression of the considered effect by the considered cause is independent from the cause in the true causal direction, contrary to the anticausal direction.

We apply SAM on this dataset composed of 1000 sampled points over 128 independent executions for various numbers of hidden units, and take note of the fit losses measured as the averaged discriminator output values. We used the experimental setting of SAM detailed in Section 6. The Figure 14 represents the evolution of the fit losses for numbers of hidden units ranging from 2 to 100 in causal and anticausal direcction. When enough capacity is given to the generators, we observe that $S_{X \to Y} < S_{Y \to X}$, allowing to recover the true causal DAG $X \to Y$.

D. Detail of the experimental results

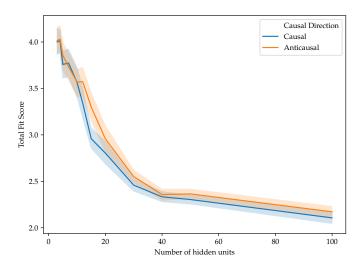


Figure 14: **Resulting losses for the causal pair plotted on Fig.13:** Evolution of the fitness loss with the number of hidden units in the generators for the two possible causal structures. For a given number of hidden units in the generators, a better (lower) loss for the causal direction indicates the ability of SAM to leverage distributional asymmetries.

Table 2: Comprehensive comparison of SAM with baseline methods on artificial data (small graphs; 20 var.). Average Precision (std. dev.) of all the algorithms for the orientation of the six types of artificial graphs with 20 variables. A higher value means a better score. A t-test is used to assess if the set of score for the best method is significantly different than the scores obtain for the other algorithms with a p-value below 0.001 (underlined values). Computational time in second required for one graph on CPU (GPU) is displayed in the last column. This table highlights the robust performance of SAM on all types of mechanisms, where it achieves the best performance except for 'GP AM' and 'Linear' mechanisms where methods that specifically search for these types of mechanisms obtain better scores.

AP	Linear	GP AM	GP Mix	Sigmoid AM	Sigmoid Mix	NN	Global	Time in s.
PC-Gauss	0.33 (0.06)	0.31 (0.07)	0.25 (0.10)	0.45 (0.10)	0.23 (0.06)	0.23 (0.05)	0.30 (0.07)	1
PC-HSIC	0.31 (0.07)	0.40 (0.06)	0.37 (0.07)	0.42 (0.08)	0.28 (0.08)	0.27 (0.03)	0.34 (0.06)	46 523
PC-RCOT	0.30 (0.04)	0.42 (0.06)	0.34 (0.06)	0.36 (0.05)	0.25 (0.05)	0.27 (0.03)	0.32 (0.05)	356
PC-RCIT	0.31(0.05)	0.37 (0.04)	0.31 (0.05)	0.35 (0.05)	0.25 (0.06)	0.27 (0.04)	0.30 (0.04)	181
GES	0.52 (0.07)	0.27 (0.06)	0.27 (0.07)	0.44 (0.14)	0.26 (0.11)	0.23 (0.07)	0.33 (0.10)	1
GIES	0.50 (0.09)	0.28 (0.07)	0.27 (0.10)	0.46 (0.14)	0.27 (0.10)	0.26 (0.09)	0.34 (0.12)	1
MMHC	0.29 (0.05)	0.23 (0.03)	0.20 (0.04)	0.31 (0.03)	0.23 (0.03)	0.25 (0.03)	0.25 (0.03)	1
LiNGAM	0.37 (0.05)	0.13 (0.03)	0.11 (0.02)	0.15 (0.06)	0.12 (0.02)	0.13 (0.03)	0.17 (0.03)	2
CAM	0.23 (0.07)	0.80 (0.07)	0.64 (0.12)	0.55 (0.11)	0.19 (0.04)	0.31 (0.10)	0.45 (0.08)	2 880
CCDr	0.33 (0.06)	0.31 (0.07)	0.25 (0.09)	0.45 (0.10)	0.23 (0.06)	0.23 (0.05)	0.30 (0.07)	2
GENIE3	0.27 (0.05)	0.47 (0.04)	0.46 (0.08)	0.40 (0.05)	0.24 (0.02)	0.31 (0.04)	0.36 (0.05)	54
SAM-lin-mse	0.31 (0.04)	0.29 (0.05)	0.29 (0.05)	0.32 (0.06)	0.28 (0.04)	0.32 (0.08)	0.30 (0.07)	332 (70)
SAM-mse	0.29 (0.04)	0.43 (0.04)	0.46 (0.10)	0.40 (0.08)	0.26 (0.05)	0.33 (0.07)	0.36 (0.05)	2 984 (91)
SAM-lin	0.49 (0.10)	0.28 (0.04)	0.29 (0.03)	0.41 (0.09)	0.35 (0.08)	0.34 (0.08)	0.30 (0.07)	14 812 (645
SAM	0.39 (0.08)	0.67 (0.08)	0.74 (0.12)	0.58 (0.13)	0.53 (0.06)	0.45 (0.09)	0.56 (0.09)	17 388 (67

SHD	Linear	GP AM	GP Mix	Sigmoid AM	Sigmoid Mix	NN
PC-Gauss	42.80 (6.74)	46.65 (4.68)	45.60 (5.45)	38.95 (9.93)	52.15 (6.46)	48.35 (7.37)
PC-HSIC	43.15 (5.04)	42.85 (7.05)	40.65 (5.16)	41.05 (9.23)	47.35 (9.32)	44.85 (6.83)
PC-RCOT	42.40 (4.42)	40.65 (6.16)	40.40 (6.38)	42.90 (8.52)	46.35 (7.49)	43.30 (6.68)
PC-RCIT	42.35 (5.09)	44.05 (5.85)	41.00 (6.24)	42.70 (9.41)	46.45 (6.37)	42.80 (7.05)
GES	43.05 (18.5)	72.20 (9.60)	57.45 (8.21)	46.55 (15.9)	75.60 (16.8)	78.05 (17.5)
GIES	42.70 (17.7)	70.45 (8.64)	57.65 (10.1)	47.55 (15.0)	57.65 (10.1)	75.25 (15.0)
MMHC	45.5 (5.25)	62.3 (4.67)	64.0 (6.85)	54.80 (9.59)	56.3 (7.16)	50.30 (7.36)
LiNGAM	36.50 (4.99)	46.70 (5.23)	43.20 (6.80)	45.80 (8.72)	52.10 (5.82)	54.80 (10.2)
CAM	71.15 (6.47)	26.80 (6.68)	42.65 (10.2)	50.90 (9.63)	75.45 (11.5)	70.50 (10.1)
CCDr	42.80 (6.40)	46.65 (4.44)	45.60 (5.17)	38.90 (9.42)	52.15 (6.12)	48.35 (6.99)
GENIE3	40.3 (6.96)	43.7 (5.81)	38.9 (7.14)	44.5 (8.41)	42.4 (5.80)	40.9 (7.23)
SAM-lin-mse	43,00 (7.29)	47.56 (6.70)	41.56 (6.31)	48.22 (9.61)	45.44 (5.56)	42.89 (7.68)
SAM-mse	46.78 (6.03)	41.00 (5.42)	36.11 (3.93)	44.33 (11.6)	49.56 (4.69)	44.89 (7.43)
SAM-lin	39.00 (6.46)	54.33 (6.29)	46.11 (4.25)	45.33 (8.86)	47.11 (6.37)	44.56 (8.69)
SAM	45.40 (5.32)	31.90 (8.53)	25.20 (4.54)	40.10 (11.7)	39.00 (4.40)	40.80 (6.05)

Table 3: Average Structural Hamming Distance (std. dev.) of all the algorithms for the orientation of the six types of artificial graphs with 20 variables. A lower value means a better score.

Table 4: Comprehensive comparison of SAM with baseline methods on artificial data (large graphs; 100 var.). Average Precision (std. dev.) of all the algorithms for the orientation of the six types of artificial graphs with 100 variables. A higher value means a better score. A t-test is used to assess if the set of score for the best method is significantly different than the scores obtain for the other algorithms with a p-value below 0.001 (underlined values). Computational time in seconds required for one graph on CPU (GPU). PC-HSIC algorithm results are not displayed as it exceeded the time limit. SAM still presents good performance on much bigger datasets without adding significant computational time, which proves the scalability of the approach. On datasets relying on Gaussian processes, CAM tops the leaderboard by a significant margin as it tries to fit Gaussian processes with the data, matching the mechanisms used.

AP	Linear	GP AM	GP Mix	Sigmoid AM	Sigmoid Mix	NN	Global	Time in s.
PC-Gauss	0.20 (0.03)	0.24 (0.03)	0.23 (0.03)	0.54 (0.04)	0.21 (0.04)	0.19 (0.03)	0.27 (0.03)	13
PC-HSIC	-	-	-	-	-	-	-	-
PC-RCOT	0.20 (0.03)	0.24 (0.03)	0.23 (0.02)	0.54 (0.04)	0.21 (0.04)	0.19 (0.03)	0.27 (0.03)	31 320
PC-RCIT	0.17 (0.03)	0.27 (0.03)	0.21 (0.02)	0.36 (0.03)	0.19 (0.02)	0.17 (0.01)	0.23 (0.02)	46 440
GES	0.38 (0.08)	0.28 (0.05)	0.23 (0.02)	0.58 (0.06)	0.37 (0.06)	0.34 (0.06)	0.36 (0.05)	1
GIES	0.38 (0.08)	0.27 (0.05)	0.23 (0.03)	0.59 (0.04)	0.38 (0.07)	0.33 (0.06)	0.36 (0.05)	5
MMHC	0.18 (0.02)	0.16 (0.01)	0.17 (0.01)	0.31 (0.02)	0.17 (0.02)	0.16 (0.01)	0.19 (0.01)	5
LiNGAM	0.22 (0.05)	0.03 (0.01)	0.03 (0.01)	0.07 (0.02)	0.05 (0.01)	0.07 (0.01)	0.08 (0.02)	5
CAM	0.28 (0.05)	0.90 (0.03)	0.66 (0.03)	0.66 (0.03)	0.22 (0.03)	0.31 (0.04)	0.50 (0.03)	45 899
CCDr	0.20 (0.03)	0.24 (0.03)	0.23 (0.02)	0.54 (0.04)	0.21 (0.04)	0.19 (0.03)	0.27 (0.03)	3
GENIE3	0.14 (0.02)	0.39 (0.02)	0.34 (0.02)	0.32 (0.02)	0.18 (0.02)	0.19 (0.01)	0.26 (0.02)	511
SAM-lin-mse	0.15 (0.02)	0.14 (0.01)	0.09 (0.01)	0.16 (0.03)	0.16 (0.02)	0.18 (0.02)	0.15 (0.02)	3 076 (74)
SAM-mse	0.15 (0.02)	0.25 (0.02)	0.11 (0.02)	0.18 (0.02)	0.18 (0.02)	0.19 (0.01)	0.18 (0.02)	18 180 (118)
SAM-lin	0.51 (0.09)	0.29 (0.04)	0.18 (0.01)	0.51 (0.04)	0.50 (0.04)	0.44 (0.07)	0.41 (0.02)	24 844 (1 980)
SAM	<u>0.53</u> (0.08)	0.58 (0.04)	0.46 (0.05)	0.63 (0.04)	<u>0.60</u> (0.07)	<u>0.45</u> (0.09)	<u>0.54</u> (0.06)	24 844 (2 041)

Table 5: Average Structural Hamming Distance (std. dev.) of all the algorithms for the orientation of the six types of artificial graphs with 100 variables. A lower value means a better score.

SHD	Linear	GP AM	GP Mix	Sigmoid AM	Sigmoid Mix	NN
PC-Gauss	262.65 (19.87)	255.35 (12.99)	250.00 (10.85)	170.55 (12.05)	258.30 (16.49)	260.80 (15.79)
PC-HSIC	-	-	-	-	-	-
PC-RCOT	262.65 (19.87)	255.35 (12.99)	250.00 (10.85)	170.55 (12.05)	258.30 (16.49)	260.80 (15.79)
PC-RCIT	253.05 (18.87)	246.30 (17.58)	246.95 (9.950)	208.75 (16.11)	244.80 (17.30)	246.05 (10.00)
GES	292.10 (38.00)	412.40 (31.04)	326.15 (17.91)	206.30 (21.39)	365.85 (32.54)	391.95 (43.10)
GIES	288.40 (34.29)	417.00 (30.76)	322.10 (18.24)	202.95 (15.75)	371.45 (29.28)	385.75 (42.37)
MMHC	275.12 (13.54)	372.41 (18.6)	345.15 (15.2)	296.51 (15.3)	315.01 (12.7)	284.93 (14.05)
LiNGAM	230.00 (12.11)	251.00 (21.76)	252.00 (10.85)	241.10 (16.78)	251.44 (17.42)	250.60 (15.69)
CAM	309.25 (26.91)	94.60 (11.20)	170.70 (11.99)	159.85 (12.39)	354.25 (18.32)	333.20 (28.84)
CCDr	262.65 (19.87)	255.35 (12.99)	250.00 (10.85)	170.55 (12.05)	258.30 (16.49)	260.80 (15.79)
GENIE3	240.2 (17.62)	252.4 (18.33)	247.0 (10.66)	238.5 (19.46)	238.3 (16.66)	237.3 (13.16)
SAM-lin-mse	238.56 (16.84)	256.78 (12.02)	247.89 (10.28)	239.67 (19.10)	238.44 (16.65)	234.11 (8.45)
SAM-mse	269.89 (20.82)	238.89 (13.08)	249.67 (11.01)	238.33 (17.57)	256.89 (19.83)	243.67 (11.02)
SAM-lin	193.89 (24.94)	251.89 (13.05)	265.67 (11.41)	196.78 (12.53)	195.67 (14.26)	199.78 (20.71)
SAM	182.30 (26.38)	186.10 (13.05)	211.60 (19.22)	158.00 (17.74)	167.60 (17.40)	186.89 (18.96)

Table 6: Average Precision (std. dev.) results for the orientation of 20 artificial graphs generated with the SynTReN simulator with 20 nodes (left) and 100 nodes (right). A t-test is used to assess if the set of score for the best method is significantly different than the scores obtain for the other algorithms with a p-value below 0.001 (underlined values). A higher value means a better score. SAM clearly outperforms the other algorithms on these datasets coming from the SynTREN generator.

AP	SynTREN 20 nodes	SynTREN 100 nodes
PC-Gauss	0.16 (0.06)	0.06 (0.01)
PC-HSIC	0.06 (0.01)	-
PC-RCOT	0.16 (0.05)	0.07 (0.02)
PC-RCIT	0.16 (0.05)	0.07 (0.01)
GES	0.14 (0.06)	0.06 (0.01)
GIES	0.12 (0.04)	0.06 (0.01)
MMHC	0.14 (0.05)	0.07 (0.01)
LiNGAM	-	-
CAM	0.21 (0.08)	0.19 (0.04)
CCDr	0.18 (0.12)	0.21 (0.05)
GENIE3	0.23 (0.07)	0.13 (0.02)
SAM-lin-mse	0.19 (0.08)	0.09 (0.02)
SAM-mse	0.40 (0.14)	0.17 (0.02)
SAM-lin	0.24 (0.23)	0.13 (0.03)
SAM	0.59 (0.15)	0.35 (0.06)

Table 7: Average Precision (AP) and Structural Hamming distance (SHD) results for the orientation of the real protein network. In terms of AP a higher value means a better score and in terms of SHD a lower value means a better score. SAM manages to obtain the best results by a significant margin.

AP	Cyto (AUPR)	Cyto (SHD)
PC-Gauss	0.16	28
PC-HSIC	-	-
PC-RCOT	0.39	22
PC-RCIT	0.41	23
GES	0.14	38
GIES	0.22	41
MMHC	0.25	23
LiNGAM	0.16	23
CAM	0.28	28
CCDr	0.22	35
GENIE3	0.32	20
SAM-lin-mse	0.26	19
SAM-mse	0.28	22
SAM-lin	0.23	20
SAM	0.45	14

Table 8: Average (std. dev.) Structural Hamming distance results for the orientation of 20 artificial graphs generated with the SynTReN simulator with 20 nodes (left), 100 nodes (middle), and real protein network (right). A lower value means a better score.

SHD	SynTREN 20 nodes	SynTREN 100 nodes	Cyto
PC-Gauss	53.42 (6.13)	262.65 (19.87)	28
PC-HSIC	24.13 (4.08)	-	-
PC-RCOT	34.21 (7.99)	213.51 (8.60)	22
PC-RCIT	33.20 (7.54)	204.95 (8.77)	23
GES	67.26 (12.26)	436.02 (18.99)	38
GIES	69.31 (12.55)	430.55 (22.80)	41
MMHC	67.2 (8.42)	346 (14.44)	38
LiNGAM	-	-	23
CAM	57.85 (9.10)	222.9 (12.38)	28
CCDr	54.97 (16.68)	228.8 (21.15)	35
GENIE3	23.6 (4.14)	153.2 (4.59)	20
SAM-lin-mse	25.44 (4.97)	240.1 (3.92)	19
SAM-mse	25.67 (6.96)	173.78 (6.36)	22
SAM-lin	30.45 (8.09)	168.89 (5.63)	20
SAM	19.02 (5.83)	160.21 (13.03)	14

Table 9: Average Precision (std. dev.) results for the orientation of 5 artificial graphs of the Dream4 In Silico Multifactorial Challenge. A higher value means a better score. GENIE3 achieves the best performance on 4 datasets, with SAM close second.

AP	NET1	NET2	NET3	NET4	NET5
PC-Gauss	0.113	0.072	0.144	0.130	0.136
PC-HSIC	0.116	0.070	0.151	0.121	0.127
PC-RCOT	0.094	0.054	0.113	0.097	0.079
PC-RCIT	0.084	0.046	0.104	0.083	0.086
GES	0.051	0.053	0.061	0.080	0.081
GIES	0.047	0.062	0.065	0.076	0.073
MMHC	0.116	0.073	0.148	0.133	0.141
LiNGAM	-	-	-	-	-
CAM	0.116	0.080	0.210	0.147	0.121
CCDr	0.088	0.099	0.114	0.119	0.165
GENIE3	0.154	0.155	0.231	0.208	0.197
SAM-lin-mse	0.108	0.136	0.204	0.159	0.111
SAM-mse	0.095	0.066	0.188	0.145	0.136
SAM-lin	0.080	0.077	0.190	0.170	0.134
SAM	0.133	0.129	0.222	0.200	0.210

Table 10: Structural Hamming distance results for the orientation of 5 artificial graphs of the Dream4 challenge generated with the GeneNetWeaver simulator with 100 nodes. A lower value means a better score.

AP	NET1	NET2	NET3	NET4	NET
PC-Gauss	183	261	200	223	203
PC-HSIC	170	249	193	210	192
PC-RCOT	174	248	193	211	191
PC-RCIT	172	248	193	211	191
GES	252	333	279	286	266
GIES	261	314	281	304	274
MMHC	188	263	206	223	203
LiNGAM	-	-	-	-	-
CAM	178	250	182	213	196
CCDr	187	248	209	227	189
GENIE3	172	245	190	208	193
SAM-lin-mse	176	249	195	211	193
SAM-mse	171	253	197	211	192
SAM-lin	175	249	190	204	191
SAM	176	251	191	209	192