CAUSAL GENERATIVE NEURAL NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

We introduce CGNN, a framework to learn functional causal models as generative neural networks. These networks are trained using backpropagation to minimize the maximum mean discrepancy to the observed data. Unlike previous approaches, CGNN leverages both conditional independences and distributional asymmetries to seamlessly discover bivariate and multivariate causal structures, with or without hidden variables. CGNN does not only estimate the causal structure, but a full and differentiable generative model of the data. Throughout an extensive variety of experiments, we illustrate the competitive results of CGNN w.r.t state-of-the-art alternatives in observational causal discovery on both simulated and real data, in the tasks of cause-effect inference, v-structure identification, and multivariate causal discovery.

1 INTRODUCTION

Deep learning models have shown extraordinary predictive abilities, breaking records in image classification (Krizhevsky et al., 2012), speech recognition (Hinton et al., 2012), language translation (Cho et al., 2014), and reinforcement learning (Silver et al., 2016). However, the predictive focus of black-box deep learning models leaves little room for explanatory power. In particular, current machine learning paradigms offer no protection to avoid mistaking correlation by causation. For example, consider that we are interested in predicting a target variable Y given a feature vector (X_1, X_2). Assume that the generative process underlying (X_1, X_2, Y) is described by the equations:

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

where (E_{X_2}, E_y) are additive noise variables. These equations tell that the values of Y are computed as a function of the values of X_1, and that the values of X_2 are computed as a function of the values of Y. The "assignment arrows" emphasize the asymmetric relations between the three random variables: we say that "X_1 causes Y", and that "Y causes X_2". However, since X_2 provides a stronger signal-to-noise ratio for the prediction of Y, the least-squares solution to this problem is $\hat{Y} = 0.25X_1 + 0.5X_2$, a typical case of inverse regression (Goldberger, 1984). Such least-squares prediction would *explain* some changes in Y as a function of changes in X_2. This is a wrong explanation, since X_2 does not cause the computation of Y. Even though there exists the necessary machinery to detect all the cause-effect relations in this example (Hoyer et al., 2009), common machine learning solutions will misunderstand how manipulating the values and distributions of (X_1, X_2), or how changing the mapping from Y to X_2, affect the values of Y. Mistaking correlation by causation can be catastrophic for agents who must plan, reason, and decide based on observation. Thus, discovering causal structures is of crucial importance.

The gold standard to discover causal relations is to perform experiments (Pearl, 2003). However, experiments are in many cases expensive, unethical, or impossible to realize. In these situations, there is a need for *observational causal discovery*, that is, the estimation of causal relations from observation alone (Spirtes et al., 2000; Peters et al., 2017). The literature in observational causal discovery is vast (see Appendix B for a brief survey), but lacks a unified solution. For instance, some approaches rely on distributional asymmetries to discover bivariate causal relations (Hoyer et al., 2009; Zhang & Hyvärinen, 2009; Daniusis et al., 2012; Stegle et al., 2010; Lopez-Paz et al., 2015; Fonollosa, 2016), while others rely on conditional independence to discover structures on three or more variables (Spirtes et al., 2000; Chickering, 2002). Furthermore, different algorithms

rely on different but equally strong assumptions, such as linearity (Shimizu et al., 2006), additive noise (Hoyer et al., 2009; Zhang & Hyvärinen, 2009), determinism (Daniusis et al., 2012), or a large corpus of annotated causal relations (Lopez-Paz et al., 2015; Fonollosa, 2016). Perhaps the most promising of the current approaches are score-based methods (Chickering, 2002). However, score-based methods rely on an external *score-function* that must be powerful enough to detect diverse causal relations. Moreover, most score-based methods only estimate the causal structure, ignoring the full generative model of the data. Finally, most of these methods are not differentiable, and thus unsuited for representation learning pipelines.

Contribution   We propose Causal Generative Neural Networks (CGNN), a framework to model functional causal models (Section 2) as generative neural networks, trained to minimize the Maximum Mean Discrepancy (MMD) to the observed data (Section 3). CGNN is a unified solution to learn causal models from data that leverages the representational power of deep generative models. By accounting for both distributional asymmetries and conditional independences, CGNN works seamlessly on bivariate and multivariate data, and is able to deal with the existence of hidden variables (confounders). CGNN uses generative networks to estimate not only the causal graph underlying the data, but a generative model. Unlike previous approaches, the generative networks used in CGNN allow non-additive noise terms to model flexible conditional distributions. Being fully differentiable, CGNN models can be trained using backpropagation, and embedded in deep learning architectures.

CGNN yields competitive results w.r.t state-of-the-art alternatives in a large variety of experiments (Section 4), including cause-effect inference (Section 4.1), v-structure identification (Section 4.2), and multivariate causal structure discovery (Section 4.3) with hidden variables (Section 4.4). Section 5 concludes with some perspectives for future work. Our code and data are available at `http://anonymous`.

2   THE LANGUAGE OF CAUSALITY: FUNCTIONAL CAUSAL MODELS

Our goal is to discover the Functional Causal Model (FCM) underlying a random variable vector $X = (X_1, \ldots, X_d)$. Formally, an FCM is a triplet $C = (\mathcal{G}, f, Q)$ producing the set of equations

$$X_i \leftarrow f_i(X_{\mathrm{Pa}(i;\mathcal{G})}, E_i), E_i \sim Q,$$

for all $i = 1, \ldots, d$. In the previous equations, each X_i is one *observed variable*, with values computed by applying a deterministic function f_i with arguments $X_{\mathrm{Pa}(i;\mathcal{G})} \subset \{X_1, \ldots, X_d\}$ and E_i, where E_i is a noise variable accounting for uncertainty.

In the language of causality, each of the previous equations expresses a direct causal relation from the causes $X_{\mathrm{Pa}(i;\mathcal{G})}$ and noise E_i to the effect X_i, described by some *causal mechanism* f_i. More specifically, we assume the existence of a Directed Acyclic Graph (DAG) $\mathcal{G}$ of d nodes where, abusing notation, each node is associated to one of the observed variables $X_1, \ldots, X_d$. Following standard graph theory notation, the set $\mathrm{Pa}(i;\mathcal{G}) \subset \{1, \ldots, d\}$ contains the indices of the parents of the node X_i in the graph. We say that there exists a direct causal relation from X_i to X_j, written $X_i \rightarrow X_j$, if there exists a directed edge from X_i to X_j in $\mathcal{G}$. Figure 1 illustrates one such FCM.

$$E_i \sim Q(E)$$
$$X_1 = f_1(E_1)$$
$$X_2 = f_2(X_1, E_2)$$
$$X_3 = f_3(X_1, E_3)$$
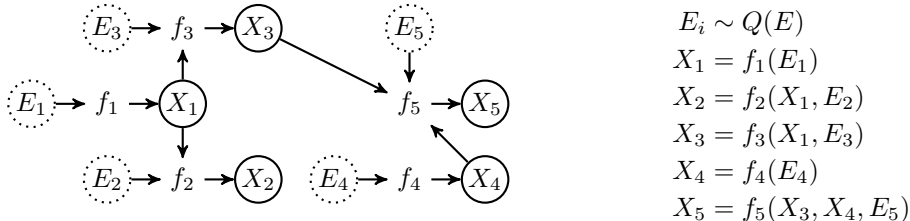$$X_4 = f_4(E_4)$$
$$X_5 = f_5(X_3, X_4, E_5)$$

Figure 1: Example of causal graph and associated functional model for $X = (X_1, \ldots, X_5)$.

FCMs are generative models.   We can draw a sample $x = (x_1, \ldots, x_d)$ from the distribution $P := P(X)$ by observing the FCM at play. First, draw $e_i \sim Q$ for all $i = 1, \ldots, d$. Second, construct $x_i = f_i(x_{\mathrm{Pa}(i;\mathcal{G})}, e_i)$ in the topological order of $\mathcal{G}$. Since this process observes but does not manipulate

the equations of the FCM, we call x one *observational sample* from P, the *observational distribution* of X. However, one FCM contains more information than the observational distribution alone, since we can decide to manipulate any of its equations and obtain a new distribution. For instance, we could decide to *set and hold constant* $X_j = 0.1$, hereby removing all the causal influences $X_k \to X_j$, for all $k \in \mathrm{Pa}(j; \mathcal{G})$. We denote by $P_{\mathrm{do}(X_j=0.1)}(X)$ the corresponding *interventional distribution*. Importantly, intervening is different from conditioning (*correlation does not imply causation*). Understanding the effect of interventions requires the (partial) knowledge of the FCM. This is why this work focuses on discovering such causal structures from data.

Formal definitions and assumptions   Two random variables (X, Y) are *conditionally independent* given Z if $P(X, Y|Z) = P(X|Z)P(Y|Z)$. Three of random variables (X, Y, Z) form a v-structure iff their causal structure is $X \to Z \leftarrow Y$. The random variable Z is a confounder (or common cause) of the pair of random variables (X, Y) if (X, Y, Z) have causal structure $X \leftarrow Z \to Y$. The skeleton $\mathcal{U}$ of a DAG $\mathcal{G}$ is obtained by replacing all the directed edges in $\mathcal{G}$ by undirected edges.

Discovering the causal structure of a random vector is a difficult task when considered in full generality. Because of this reason, the literature in causal inference relies on a set of common assumptions (Pearl, 2003). The *causal sufficiency* assumption states that there are no unobserved confounders. The *causal Markov* assumption states that all the d-separations in the causal graph $\mathcal{G}$ imply conditional independences in the observational distribution P. The *causal faithfulness* assumption states that all the conditional independences in the observational distribution P imply d-separations in the causal graph $\mathcal{G}$. We call *Markov equivalence class* to the set of graphs containing the same set of d-separations. When using the causal faithfulness assumption and conditional independence information, we are able to recover the Markov equivalence class of the causal structure underlying a random vector –which, in some cases contains one graph, the causal structure itself. Markov equivalence classes are DAGs where some of the edges remain undirected.

Learning FCMs from data using score methods   Consider a random vector $X = (X_1, \ldots, X_d)$ following the FCM $C = (\mathcal{G}, f, Q)$ with associated observational distribution P. Furthermore, assume access to n samples drawn from P, denoted by $\mathcal{D} = \{x_i\}_{i=1}^n$, where $x_i = (x_{i,1}, \ldots, x_{i,d}) \in \mathbb{R}^d$ for all $i = 1, \ldots, n$. Given these data, the goal of *observational causal discovery* is to estimate the underlying causal DAG $\mathcal{G}$ and the causal mechanisms f.

One family of methods for observational causal discovery are *score-based methods* (Chickering, 2002). In essence, score-based methods rely on some *score-function* $S(\mathcal{G}, \mathcal{D})$ to measure the fit between a candidate set $\{\mathcal{G}, f\}$ and the observed data $\mathcal{D}$. Then, we select the DAG on d variables achieving the maximum score as measured by S. As an example of score-function, consider the Bayesian Information Criterion (BIC): $S(\mathcal{G}, \mathcal{D}) = \sum_{i=1}^n \sum_{j=1}^d \log p_{\hat{\theta}_j}(x_{i,j}|x_{i,\mathrm{Pa}(j;\mathcal{G})}) - \lambda|\mathcal{G}|$, where $p_{\hat{\theta}_j}$ the maximum-likelihood estimate of a simple parametric family of conditional distributions $p_{\theta \in \Theta}$ allowing efficient density evaluation. The term $\lambda \in [0, \infty)$ penalizes the number of edges (that is, the model complexity assuming equal number of parameters per edge) in the graph. Finally, we may associate each edge $X_i \to X_j$ in $\mathcal{G}$ to an importance or confidence score proportional to its contribution to the overal loss: as $V_{X_i \to X_j} = S(\mathcal{G}, \mathcal{D}) - S(\mathcal{G} - \{X_i \to X_j\}, \mathcal{D})$.

A naïve score-based method would enumerate all the DAGs of d variables and select the one maximizing S. Unfortunately, the number of DAGs over d nodes is super-exponential in d. Thus, the brute-force search of the best DAG is intractable, even for moderate d. Inspired by Tsamardinos et al. (2006); Nandy et al. (2015), we assume in this paper known graph skeletons. Such a skeleton may arise from expert knowledge or a feature selection algorithm algorithm (Yamada et al., 2014) under standard assumptions such as causal Markov, faithfulness, and sufficiency. Given a skeleton with k edges, causal discovery reduces to selecting one out of the $O(2^k)$ possible edge orientations.

3   CAUSAL GENERATIVE NEURAL NETWORKS

This section proposes a framework to learn FCMs from data by leveraging the representational power of generative neural networks. In particular, we propose to estimate FCMs C as $\hat{C} = (\hat{\mathcal{G}}, \hat{f}, \hat{Q})$, with

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

for all $i = 1, \ldots, d$. Here, $\hat{\mathcal{G}}$ is the estimated causal graph of X, the functions $\hat{f} = (\hat{f}_1, \ldots, \hat{f}_d)$ are the estimated causal mechanisms of X producing the estimated observed variables $\hat{X} = (\hat{X}_1, \ldots, \hat{X}_d)$, and the estimated noise variables $\hat{E} = (\hat{E}_1, \ldots, \hat{E}_d)$ are sampled from a fixed distribution $\hat{Q}$. Given the estimated FCM (1), we can draw n samples from its observational distribution $\hat{P}$ (see Section 2) and construct the *estimated observational samples* $\hat{\mathcal{D}} = \{\hat{x}_i\}_{i=1}^n$.

We parametrize the equations (1) as generative neural networks, also known as conditional generators (Goodfellow et al., 2014). Without loss of generality, we assume that the independent noise variables $\hat{E}$ are sampled from an univariate Normal distribution (Stegle et al., 2010). Then, we propose the following score-function to measure the fit between a candidate structure $\hat{\mathcal{G}}$ and data $\mathcal{D}$:

$$S(\hat{\mathcal{G}}, \mathcal{D}) = -\widehat{\mathrm{MMD}}_k(\mathcal{D}, \hat{\mathcal{D}}) - \lambda|\hat{\mathcal{G}}|. \tag{2}$$

where MMD is the Maximum Mean Discrepancy statistic (Gretton et al., 2007):

$$\widehat{\mathrm{MMD}}_k(\mathcal{D}, \hat{\mathcal{D}}) = \left\| \mu_k(\mathcal{D}) - \mu_k(\hat{\mathcal{D}}) \right\|_{\mathcal{H}_k} = \frac{1}{n^2} \sum_{i,j}^n k(x_i, x_j) + \frac{1}{n^2} \sum_{i,j}^n k(\hat{x}_i, \hat{x}_j) - \frac{2}{n^2} \sum_{i,j}^n k(x_i, \hat{x}_j).$$
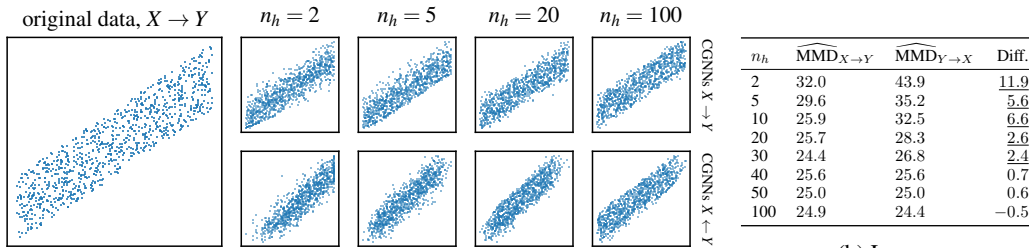
The MMD statistic scores a graph $\hat{\mathcal{G}}$ by measuring the discrepancy between the *data* observational distribution P and the *estimated* observational distribution $\hat{P}$, on the basis of their samples. When using a characteristic kernel k such as the Gaussian kernel $k(x, x') = \exp(-\gamma \|x - x'\|_2^2)$, MMD is an well-defined score-function: it is zero if and only if $P = \hat{P}$ as $n \to \infty$ (Gretton et al., 2007). Since the computation of $\widehat{\mathrm{MMD}}_k$ takes $O(n^2)$ time, our experiments will also consider an approximation based on m random features (Lopez-Paz, 2016), denoted by $\widehat{\mathrm{MMD}}_k^m$. Appendix A offers a brief exposition on MMD. In a nutshell, CGNN implements Occam's razor to prefer simpler models as causal. Unlike previous methods, CGNN can seamlessly leverage both distributional asymmetries (due to the representational power of generative networks) and conditional independences (due to the joint minimization of those networks using MMD) to score both bivariate and multivariate graphs.

For a differentiable kernel k such as the Gaussian kernel, the score function (2) is differentiable and therefore CGNN is trainable using backpropagation. CGNN is a directed acyclic graph of conditional generator networks that result in a flexible generative model of the data causal structure.

Searching causal graphs with CGNN Using the CGNN score (2), we propose the following greedy approach to orient a given skeleton:

1. Orient each $X_i - X_j$ as $X_i \to X_j$ or $X_j \to X_i$ by selecting the 2-variable CGNN with the best score.

2. Remove all cycles: all paths starting from a random set of nodes are followed iteratively until all nodes are reached; an edge pointing towards an already visited node forms a cycle, so is reversed.

3. For a number of iterations, reverse the edge that leads to the maximum improvement over a d-variable CGNN, without creating a cycle .

Dealing with hidden confounders The search method above assumes the causal sufficiency assumptions: or, the non-existence of hidden confounders. We address this issue in a variant of our algorithm as follows. When assuming the existence of confounders, each edge $X_i - X_j$ in the skeleton is due to one out of three possibilities: either $X_i \to X_j$, $X_j \leftarrow X_i$, or there exists an unobserved variable $E_{i,j}$ such that $X_i \leftarrow E_{i,j} \to X_j$. Therefore, each equation in the FCM is extended to: $X_i \leftarrow f_i(X_{\mathrm{Pa}(i;\mathcal{G})}, E_{i,\mathrm{Ne}(i;\mathcal{S})}, E_i)$, where $\mathrm{Ne}(i;\mathcal{S}) \subset \{1, \ldots d\}$ is the set of indices of the variables adjacent to X_i in the skeleton. Here each $E_{i,j} \sim Q$ and denotes the hypothetical unobserved common causes of X_i and X_j. For instance, if we hide X_1 from the FCM described in Figure 1, this would require considering a confounder $E_{2,3}$. Finally, when considering hidden confounders, the third step above considers three possible mutations of the graph: reverse, add, or remove an edge. Here, the term $\lambda|\hat{\mathcal{G}}|$ takes an active role and promotes simple graphs.

original data, $X \rightarrow Y$	$n_h = 2$	$n_h = 5$	$n_h = 20$	$n_h = 100$

(a) Samples.

n_h	$\widehat{\text{MMD}}_{X \rightarrow Y}$	$\widehat{\text{MMD}}_{Y \rightarrow X}$	Diff.
2	32.0	43.9	11.9
5	29.6	35.2	5.6
10	25.9	32.5	6.6
20	25.7	28.3	2.6
30	24.4	26.8	2.4
40	25.6	25.6	0.7
50	25.0	25.0	0.6
100	24.9	24.4	−0.5

(b) Losses.

Figure 2: Samples and MMDs for CGNN models of different complexities (number of neurons) modeling the causal direction $X \rightarrow Y$ (top row) and the anticausal direction $X \leftarrow Y$ (bottom row) of a simple example. MMDs are averaged over 32 runs, underlined numbers indicate statistical significance at $\alpha = 10^{-3}$.

4 EXPERIMENTS

We evaluate the performance of CGNN at discovering different types of causal structures. We study the problems of discovering cause-effect relations (Section 4.1), v-structures (Section 4.2), and multivariate causal structures without (Section 4.3) or with (Section 4.4) hidden variables.

Our experiments run at an Intel Xeon 2.7GHz CPU, and an NVIDIA 1080Ti GPU. MMD uses a sum of Gaussian kernels with bandwidths $\gamma \in \{0.005, 0.05, 0.25, 0.5, 1, 5, 50\}$. CGNN uses one-hidden-layer neural networks with n_h ReLU units, trained with the Adam optimizer (Kingma & Ba, 2014) and initial learning rate 0.01. According to preliminary experiments, using all data for both training and evaluating models produces good results, since resampling noise variables conbats overfitting. Also, our best results follow when using the whole data as a minibatch. We train CGNN during $n_{\text{train}} = 1000$ epochs and evaluate it on $n_{eval} = 500$ generated samples. We ensemble CGNN training over $n_{run} = 32$ random initializations for $\widehat{\text{MMD}}_k$ and $n_{run} = 64$ for $\widehat{\text{MMD}}_k^m$.

Regarding CGNN model selection, the number of hidden units n_h is the most sensitive hyper-parameter, and should be cross-validated for every application. The number of hidden units n_h relates to the flexibility of the CGNN to model each of the causal mechanisms. For small n_h, we may miss some of the patterns in the data. For a large n_h, we may find over-complicated explanations from effects to causes. Therefore, our interest is to find the smallest n_h explaining the data well. We illustrate such Occam's razor principle in Figure 2, where we learn two bivariate CGNNs of different complexity ($n_h = 2, 5, 20, 100$) using data from the FCM: $X \sim \text{Uniform}[-2, 2], Y \leftarrow X + \text{Uniform}[0, 0.5]$. Figure 2 shows the associated MMDs (averaged on 32 runs), confirming the importance of capacity control (Zhang & Hyvärinen, 2009). On this illustrative case the most discriminative value appears for $n_h = 2$.

4.1 DISCOVERING CAUSE-EFFECT RELATIONS

Under the causal sufficiency assumption, the statistical dependence between two random variables X and Y is due to a causal relation $X \rightarrow Y$ or due to a causal relation $X \leftarrow Y$ (Pearl, 2003). Given data from the observational distribution $P(X, Y)$, this section evaluates the performance of CGNN to decide whether $X \rightarrow Y$ or $X \leftarrow Y$.

In the following, we use five cause-effect inference datasets, covering a wide range of associations. *CE-Cha* contains 300 cause-effect pairs from the challenge of Guyon (2013). *CE-Net* contains 300 artificial cause-effect pairs generated using random distributions as causes, and neural networks as causal mechanisms. *CE-Gauss* contains 300 artificial cause-effect pairs as generated by Mooij et al. (2016), using random mixtures of Gaussians as causes, and Gaussian process priors as causal mechanisms. *CE-Multi* contains 300 artificial cause-effect pairs built with random linear and polynomial causal mechanisms. In this dataset, we simulate additive or multiplicative noise, applied before or after the causal mechanism. *CE-Tueb* contains the 99 real-world scalar cause-effect pairs from

the Tübingen dataset (Mooij et al., 2016), concerning domains such as climatology, finance, and medicine. We set $n \leq 1,500$. See our implementation for details.

CGNN is compared to following algorithms: The Additive Noise Model, or *ANM* (Mooij et al., 2016), with Gaussian process regression and HSIC independence test. The Linear Non-Gaussian Additive Model, or *LiNGAM* (Shimizu et al., 2006), a variation of Independent Component Analysis to identify linear causal relations. The Information Geometric Causal Inference, or *IGCI* (Daniusis et al., 2012), with entropy estimator and Gaussian reference measure. The Post-Non-Linear model, or *PNL* (Zhang & Hyvärinen, 2009), with HSIC test. The *GPI* method (Stegle et al., 2010), where the Gaussian process regression with higher marginal likelihood is preferred as the causal direction. The Conditional Distribution Similarity statistic, or *CDS* (Fonollosa, 2016), which prefers the causal direction with lowest variance of conditional distribution variances. The award-winning method *Jarfo* (Fonollosa, 2016), a random forest classifier trained on the ChaLearn Cause-effect pairs and hand-crafted to extract 150 features, including the methods ANM, IGCI, CDS, and LiNGAM.

The code for ANM, IGCI, PNL, GPI, and LiNGAM is available at `https://github.com/ssamot/causality`. We follow a leave-one-dataset-out scheme to select the best hyperparameters for each method. For CGNN, we search for the number of hidden neurons $n_h \in \{5, 10, 15, 20, 25, 30, 35, 40, 50, 100\}$. The leave-one-dataset-out hyperparameter selection chooses n_h equal to 35, 35, 40, 30, 40 for the *CE-Cha*, *CE-Net*, *CE-Gauss*, *CE-Multi* and *CE-Tueb* datasets respectively. For ANM, we search the Gaussian kernel bandwidth γ used in the Gaussian process regression in $\{0.01, 0.1, 0.2, 0.5, 0.8, 1, 1.2, 1.5, 2, 5, 10\}$. For LiNGAM and IGCI, there are no parameters to set. For PNL, we search for the significance level of the independence test $\alpha \in \{0.0005, 0.005, 0.01, 0.025, 0.04, 0.05, 0.06, 0.075, 0.1, 0.25, 0.5\}$. For GPI, we use the default parameters from the original implementation. For CDS, we search for the best discretization of the cause variable into $\{1, \ldots, 10\}$ levels. For Jarfo, we train the random forest using $4,000$ cause-effect pairs generated in the same way as the proposed datasets, except the one used for testing.

Table 1 reports the Area Under the Precission/Recall Curve (AUPRC) associated to the binary classification problem of deciding "$X \rightarrow Y$" or "$X \leftarrow Y$" for each cause-effect pair, for all methods and datasets. The table also shows the computational time (in both CPU and GPU), and computational complexity for methods. The least performing methods are those based on linear regression. The methods CDS and IGCI perform well on a few datasets. This indicates the existence of certain biases (such as causes having always higher entropy than effects) on such datasets. ANM performs well when the additive noise assumption holds (for instance, *CE-Gauss*), but badly otherwise. PNL, a generalization of ANM, compares favorably to these methods. Jarfo, the method using thousands of training cause-effect pairs to learn from data, performs well on artificial data but badly on real examples. The generative methods GPI and CGNN show a good performance on most datasets, including the real-world cause-effect pairs CE-Tueb. In terms of computation, generative methods are the most expensive alternatives. Fortunately for CGNN, the approximation of MMD with random features (see Appendix A) does not degrade performance, but reduces the computation time. Overall, these results suggest that CGNN is competitive compared to the state-of-the-art on the cause-effect inference problem, where it is necessary to discover distributional asymmetries.

4.2 DISCOVERING V-STRUCTURES

This section studies the performance of CGNN on the task of identifying the causal structure of three random variables (A, B, C) with skeleton $A - B - C$. The four possible structures are the *chain* $A \rightarrow B \rightarrow C$, the *reverse chain* $A \leftarrow B \leftarrow C$, the *v-structure* $A \rightarrow B \leftarrow C$, and the *reverse v-structure* $A \leftarrow B \rightarrow C$. Other skeletons are not of interest, since the absence of an edge (statistical independence) is easier to discover, and the remaining edge could be oriented using the bivariate methods described in the previous section. Three of the possible structures (the chain, the reverse chain, and the reverse v-structure) are Markov equivalent, and therefore indistinguishable from each other using statistics alone. Therefore, the goal of this section is to use CGNN to determine whether $P(A, B, C)$ follows or not an FCM with causal graph: $A \rightarrow B \leftarrow C$.

This section considers an FCM with identity causal mechanisms and Normal noise variables; for instance, $B \leftarrow A + E_B$, where $E_B \sim \mathcal{N}(0, 1)$. Therefore, the joint distribution of one cause and its effect is symmetrical (a two-dimensional Gaussian), and the bivariate methods used in the previous

Table 1: AUPRC for the cause-effect experiments. For the dataset Tueb, the weighted accuracy (Mooij et al., 2016, c.f.) is shown in parenthesis.

method	Cha	Net	Gauss	Multi	Tueb	time	complexity
Best fit	56.4	77.6	36.3	55.4	58.4 (44.9)	< 1s CPU	$O(N^3)$
LiNGAM	54.3	43.7	66.5	59.3	39.7 (44.3)	< 1s CPU	$O(N)$
CDS	55.4	89.5	84.3	37.2	59.8 (65.5)	< 1s CPU	$O(N)$
IGCI	54.4	54.7	33.2	80.7	60.7 (62.6)	< 1s CPU	$O(N)$
ANM	66.3	85.1	88.9	35.5	53.7 (59.5)	< 1s CPU	$O(N^3)$
PNL	73.1	75.5	83.0	49.0	68.1 (66.2)	23 min CPU	$O(N^2)$
Jarfo	<u>79.5</u>	<u>92.7</u>	85.3	94.6	54.5 (59.5)	< 1s CPU	$O(N^3)$
GPI	67.4	88.4	<u>89.1</u>	65.8	66.4 (62.6)	32min CPU	$O(N^3)$
CGNN ($\widehat{\textbf{MMD}}_k$)	73.6	89.6	82.9	<u>96.6</u>	<u>79.8</u> (74.4)	24 min GPU	$O(N^2)$
CGNN ($\widehat{\textbf{MMD}}_k^m$)	76.5	87.0	88.3	94.2	76.9 (72.7)	5 min GPU	$O(N)$

Table 2: CGNN MMDs of all estimated structure (rows), versus true structures (columns). The score of the v-structure (bold) discriminates this configuration from the chain structure and the reverse v-structure.

	chain	reverse v-structure	v-structure
score for chain	0.122 (0.009)	0.124 (0.007)	0.172 (0.005)
score for reverse chain	0.121 (0.006)	0.127 (0.008)	0.171 (0.004)
score for reverse v-structure	0.122 (0.007)	0.125 (0.006)	0.172 (0.004)
score for v-structure	**0.202** (0.004)	**0.180** (0.005)	**0.127** (0.005)

section all fail to apply. To succeed at this task, a causal discovery method must reason about the conditional independences between the three random variables at play.

Our protocol fits one CGNN for each of the four possible causal graphs with skeleton $A - B - C$. Then, we evaluate MMD of each of the four CGNN models, and prefer the one achieving the lowest MMD. Table 2 summarizes our results: CGNN assigns the lowest MMD to the v-structure hypothesis on those datasets generated by v-structures, and assigns the largest MMD to the v-structure hypothesis on those datasets not generated by v-structures. Sections 4.1 and 4.2 show the two complementary properties of the CGNN: leveraging distributional asymmetries and conditional independences.

4.3 DISCOVERING MULTIVARIATE CAUSAL STRUCTURES WITHOUT HIDDEN CONFOUNDERS

Consider a random vector $X = (X_1, ..., X_d)$. Our goal is to find the FCM of X under the causal Markov, faithfulness and causal sufficiency assumptions. At this point, we will assume known skeleton, so the problem reduces to orienting every edge. To that end, all experiments provide all algorithms *the true graph skeleton*, so their ability to orient edges is compared in a fair way. This allows us to separate the task of orienting the graph from that of uncovering the skeleton. We draw 500 samples from four artificial causal graphs $\mathcal{G}_2$, $\mathcal{G}_3$, $\mathcal{G}_4$, and $\mathcal{G}_5$ on 20 variables. For $i = \{2, \dots, 5\}$, the variables in the graph $\mathcal{G}_i$ have a random number of parents between 1 and i. We build the graphs with polynomial mechanisms, and additive/multiplicative noise.

We compare CGNN to the PC algorithm (Spirtes et al., 2000), the score-based method GES (Chickering, 2002), ANM, LiNGAM, and Jarfo. For PC, we employ the better-performing, order-independent version of the PC algorithm proposed by Colombo & Maathuis (2014). PC needs the specification of a conditional independence test. We compare PC-Gaussian, which employs a Gaussian conditional independence test on Fisher z-transformations, and PC-HSIC, which uses the HSIC conditional independence test with the Gamma approximation (Gretton et al., 2005). For both conditional independence tests, the significance level achieving best results is $\alpha = 0.1$. For GES, the best penalization parameter is $\lambda = 3.11$. PC and GES are implemented in the *pcalg* package (Kalisch et al., 2012). For CGNN, n_h is set to 20. We also compare to pairwise methods presented in the last section : ANM, LiNGAM, and Jarfo.

Table 3: AUPRC on multivariate causal discovery for all methods and graphs.

method	$\mathcal{G}_2$	$\mathcal{G}_3$	$\mathcal{G}_4$	$\mathcal{G}_5$	time
PC-Gaussian	82.3 ±4	80.0 ±7	88.1 ±10	92.5±4	1 to 10s CPU
PC-HSIC	93.4 ±3	93.0 ±4	98.9 ±2	96.7 ±2	2 to 15h CPU
GES	75.3 ±7	73.6 ±7	69.3±11	75.6 ±5	1 to 2s CPU
LiNGAM	64.4 ±4	71.1 ±1	71.6 ±7	72.1 ±1	1 to 2s CPU
ANM	72.9 ±9	72.5 ±4	79.9 ±5	71.8 ±2	5 to 20s CPU
Jarfo	69.9 ±9	87.3 ±3	88.5 ±5	70.2 ±7	10 to 30s CPU
CGNN ($\widehat{\mathrm{MMD}}_k^m$)	94.5 ±2	84.9 ±9	88.4 ±4	87.0 ±6	3 to 4 h GPU
CGNN ($\widehat{\mathrm{MMD}k}$)	95.7 ±1	96.5 ±3	97.2 ±3	88.2 ±6	4 to 5 h GPU

Table 4: AUPRC for experiments on causal discovery with confounders.

method	$\mathcal{G}_2$	$\mathcal{G}_3$	$\mathcal{G}_4$	$\mathcal{G}_5$	time
RFCI-Gaussian	53.4 ±11	49.0 ±10	51.1 ±13	63.6 ±11	10 to 30s CPU
RFCI-HSIC	76.2 ±11	65.2 ±7	73.1 ±6	72.9 ±6	3 to 18h CPU
Jarfo	64.1 ±11	62.1 ±7	72.2 ±7	64.9 ±4	10 to 30s CPU
CGNN ($\hat{\mathrm{MMD}}_k$)	79.8 ±12	76.1 ±11	84.4 ±7	70.9 ±4	4 to 7 h GPU

Table 3 displays the performance of all algorithms measured from the area under the precision/recall curve. Overall, the best performing method is PC-HSIC, followed closely by CGNN. The performance of PC-HSIC is best for denser graphs. This is because the PC algorithm uses a majority voting rule to decide each orientation, one strategy well suited to dense known skeletons, since one edge belongs to multiple v-structures. However, CGNN offers the advantage to orient all the edges (while some edges remain undirected by PC-HSIC) and to deliver a full generative model useful for simulation (while PC-HSIC only gives the graph). To explore the scalability of our method, we were able to extend the experiment on 5 graphs $\mathcal{G}_3$ with 100 variables, achieving an AUPRC of 85.5 ± 4, in 30 hours of computation on four NVIDIA 1080Ti GPUs.

4.4 DISCOVERING MULTIVARIATE CAUSAL STRUCTURES WITH HIDDEN CONFOUNDERS

In real applications, some confounding variables may be unobserved. We propose to use the same data from the previous section, but hide some of the 20 observed variables in the graph. More specifically, we hide three random variables that cause at least two others in the same graph. Consequently, the skeleton now includes additional edges $X - Y$ for all pairs of variables (X, Y) that are consequences of the same hidden cause (confounder). The goal in this section is to orient the edges due to direct causal relations, and to remove those edges due to confounding.

We compare CGNN to the RFCI algorithm (Colombo et al., 2012), which is a modification of the PC algorithm that accounts for hidden variables. As done in the previous section, we compare variants of RFCI based on Gaussian or HSIC (Zhang et al., 2012) conditional independence tests. We also evaluate the performance of the data-driven method Jarfo, this time trained on the whole Kaggle data of (Guyon, 2013; 2014), in order to classify relations into $X \to Y$, $X \leftarrow Y$, or $X \leftarrow Z \to Y$ (confounder). For CGNN, we penalize the objective function (2) with $\lambda = 5 \times 10^{-5}$.

Table 4 shows that CGNN is robust to the existence of hidden confounders, achieving state-of-the-art performance in this task. Interestingly, the true causal relations exhibit a high confidence score, while edges due to confounding effects are removed or have low confidence scores. Overall, CGNN performs best on the graphs $\mathcal{G}_2, \mathcal{G}_3$, and $\mathcal{G}_4$, and is slightly outperformed by RFCI-HSIC on the denser graph $\mathcal{G}_5$. However, CGNN is the only approach providing a generative model of the data.

5 Conclusion

We introduced a new framework to learn functional causal models based on generative neural networks. We train these networks by minimizing the discrepancy between their generated samples and the observed data. Such models are instances of the bigger family of FCMs for which each function is a shallow neural network with n_h hidden units.

We believe that our approach opens new avenues of research, both from the point of view of leveraging the power of deep learning in causal discovery and from the point of view of building deep networks with better structure interpretability. Once the model is learned, the CGNNs present the advantage to be fully parametrized and may be used to simulate interventions on one or more variables of the model and evaluate their impact on a set of target variables. This usage is relevant in a wide variety of domains, typically among medical and sociological domains.

Five directions for future work are to i) lower the computational cost of CGNN, ii) extend CGNN to deal with categorical data, iii) explore better heuristics for causal graph search, iv) adapt our methods for temporal data and v) obtain theoretical guarantees for basic use cases.

References

D. M. Chickering. Optimal structure identification with greedy search. *JMLR*, 2002.

K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv*, 2014.

D. Colombo and M. H. Maathuis. Order-independent constraint-based causal structure learning. *JMLR*, 2014.

D. Colombo, M. H. Maathuis, M. Kalisch, and T. S. Richardson. Learning high-dimensional directed acyclic graphs with latent and selection variables. *The Annals of Statistics*, 2012.

P. Daniusis, D. Janzing, J. Mooij, J. Zscheischler, B. Steudel, K. Zhang, and B. Schölkopf. Inferring deterministic causal relations. *arXiv*, 2012.

M. Drton and M. H. Maathuis. Structure learning in graphical modeling. *Annual Review of Statistics and Its Application*, 2017.

J. Fonollosa. Conditional distribution variability measures for causality detection. *arXiv*, 2016.

A. S. Goldberger. Reverse regression and salary discrimination. *Journal of Human Resources*, 1984.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. *NIPS*, 2014.

A. Gretton, R. Herbrich, A. Smola, O. Bousquet, and B. Schölkopf. Kernel methods for measuring independence. *JMLR*, 2005.

A. Gretton, K. M. Borgwardt, M. Rasch, B. Schölkopf, A. J. Smola, et al. A kernel method for the two-sample-problem. *NIPS*, 2007.

I. Guyon. Chalearn cause effect pairs challenge, 2013. URL http://www.causality.inf.ethz.ch/cause-effect.php.

I. Guyon. Chalearn fast causation coefficient challenge. 2014.

C. Heinze-Deml, M. H. Maathuis, and N. Meinshausen. Causal structure learning. *arXiv*, 2017.

G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 2012.

P. O. Hoyer, D. Janzing, J. M. Mooij, J. Peters, and B. Schölkopf. Nonlinear causal discovery with additive noise models. *NIPS*, 2009.

M. Kalisch and P. Bühlmann. Causal structure learning and inference: a selective review. *Quality Technology & Quantitative Management*, 2014.

M. Kalisch, M. Mächler, D. Colombo, M. H. Maathuis, P. Bühlmann, et al. Causal inference using graphical models with the r package pcalg. *Journal of Statistical Software*, 2012.

D. P. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. *ICLR*, 2014.

A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *NIPS*, 2012.

D. Lopez-Paz. *From dependence to causation*. PhD thesis, 2016.

D. Lopez-Paz and M. Oquab. Revisiting classifier two-sample tests. *ICLR*, 2017.

D. Lopez-Paz, K. Muandet, B. Schölkopf, and I. O. Tolstikhin. Towards a learning theory of cause-effect inference. *ICML*, 2015.

D. Lopez-Paz, R. Nishihara, S. Chintala, B. Schölkopf, and L. Bottou. Discovering causal signals in images. *CVPR*, 2017.

J. M. Mooij, J. Peters, D. Janzing, J. Zscheischler, and B. Schölkopf. Distinguishing cause from effect using observational data: methods and benchmarks. *JMLR*, 2016.

P. Nandy, A. Hauser, and M. H. Maathuis. High-dimensional consistency in score-based and hybrid structure learning. *arXiv*, 2015.

J. M. Ogarrio, P. Spirtes, and J. Ramsey. A hybrid causal search algorithm for latent variable models. *Conference on Probabilistic Graphical Models*, 2016.

J. Pearl. Causality: models, reasoning and inference. *Econometric Theory*, 2003.

J. Peters, D. Janzing, and B. Schölkopf. *Elements of Causal Inference - Foundations and Learning Algorithms*. MIT Press, 2017.

J. D. Ramsey. Scaling up greedy causal search for continuous variables. *arXiv*, 2015.

W. Rudin. Fourier analysis on groups, 1962.

E. Sgouritsa, D. Janzing, P. Hennig, and B. Schölkopf. Inference of cause and effect with unsupervised inverse regression. *AISTATS*, 2015.

S. Shimizu, P. O. Hoyer, A. Hyvärinen, and A. Kerminen. A linear non-gaussian acyclic model for causal discovery. *JMLR*, 2006.

D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, et al. Mastering the game of Go with deep neural networks and tree search. *Nature*, 2016.

P. Spirtes, C. Meek, T. Richardson, and C. Meek. An algorithm for causal inference in the presence of latent variables and selection bias. 1999.

P. Spirtes, C. N. Glymour, and R. Scheines. *Causation, prediction, and search*. 2000.

A. Statnikov, M. Henaff, N. I. Lytkin, and C. F. Aliferis. New methods for separating causes from effects in genomics data. *BMC genomics*, 2012.

O. Stegle, D. Janzing, K. Zhang, J. M. Mooij, and B. Schölkopf. Probabilistic latent variable models for distinguishing between cause and effect. *NIPS*, 2010.

I. Tsamardinos, L. E. Brown, and C. F. Aliferis. The max-min hill-climbing bayesian network structure learning algorithm. *Machine learning*, 2006.

M. Yamada, W. Jitkrittum, L. Sigal, E. P. Xing, and M. Sugiyama. High-dimensional feature selection by feature-wise kernelized lasso. *Neural computation*, 2014.

K. Zhang and A. Hyvärinen. On the identifiability of the post-nonlinear causal model. *UAI*, 2009.

K. Zhang, J. Peters, D. Janzing, and B. Schölkopf. Kernel-based conditional independence test and application in causal discovery. *arXiv*, 2012.

A THE MAXIMUM MEAN DISCREPANCY (MMD) STATISTIC

The Maximum Mean Discrepancy (MMD) statistic (Gretton et al., 2007) measures the distance between two probability distributions P and $\hat{P}$, defined over $\mathbb{R}^d$, as the real-valued quantity

$$\text{MMD}_k(P, \hat{P}) = \left\| \mu_k(P) - \mu_k(\hat{P}) \right\|_{\mathcal{H}_k}.$$

Here, $\mu_k = \int k(x, \cdot) \mathrm{d}P(x)$ is the *kernel mean embedding* of the distribution P, according to the real-valued symmetric kernel function $k(x, x') = \langle k(x, \cdot), k(x', \cdot) \rangle_{\mathcal{H}_k}$ with associated reproducing kernel Hilbert space $\mathcal{H}_k$. Therefore, μ_k summarizes P as the expected value of the features computed by k over samples drawn from P.

In practical applications, we do not have access to the distributions P and $\hat{P}$, but to their respective sets of samples $\mathcal{D}$ and $\hat{\mathcal{D}}$, defined in Section 3. In this case, we approximate the kernel mean embedding $\mu_k(P)$ by the *empirical kernel mean embedding* $\mu_k(\mathcal{D}) = \frac{1}{|\mathcal{D}|} \sum_{x \in \mathcal{D}} k(x, \cdot)$, and respectively for $\hat{P}$. Then, the empirical MMD statistic is

$$\widehat{\text{MMD}}_k(\mathcal{D}, \hat{\mathcal{D}}) = \left\| \mu_k(\mathcal{D}) - \mu_k(\hat{\mathcal{D}}) \right\|_{\mathcal{H}_k} = \frac{1}{n^2} \sum_{i,j}^{n} k(x_i, x_j) + \frac{1}{n^2} \sum_{i,j}^{n} k(\hat{x}_i, \hat{x}_j) - \frac{2}{n^2} \sum_{i,j}^{n} k(x_i, \hat{x}_j).$$

Importantly, the empirical MMD tends to zero as $n \to \infty$ if and only if $P = \hat{P}$, as long as k is a characteristic kernel (Gretton et al., 2007). This property makes the MMD an excellent choice to model how close the observational distribution P is to the estimated observational distribution $\hat{P}$. Throughout this paper, we will employ a particular characteristic kernel: the Gaussian kernel $k(x, x') = \exp(-\gamma \|x - x'\|_2^2)$, where $\gamma > 0$ is a hyperparameter controlling the smoothness of the features.

In terms of computation, the evaluation of $\text{MMD}_k(\mathcal{D}, \hat{\mathcal{D}})$ takes $O(n^2)$ time, which is prohibitive for large n. When using a shift-invariant kernel, such as the Gaussian kernel, one can invoke Bochner's theorem (Rudin, 1962) to obtain a linear-time approximation to the empirical MMD (Lopez-Paz et al., 2015), with form

$$\widehat{\text{MMD}}_k^m(\mathcal{D}, \hat{\mathcal{D}}) = \left\| \hat{\mu}_k(\mathcal{D}) - \hat{\mu}_k(\hat{\mathcal{D}}) \right\|_{\mathbb{R}^m}$$

and $O(mn)$ evaluation time. Here, the *approximate empirical kernel mean embedding* has form

$$\hat{\mu}_k(\mathcal{D}) = \sqrt{\frac{2}{m}} \frac{1}{|\mathcal{D}|} \sum_{x \in \mathcal{D}} \left[\cos(\langle w_1, x \rangle + b_1), \dots, \cos(\langle w_m, x \rangle + b_m) \right],$$

where w_i is drawn from the normalized Fourier transform of k, and $b_i \sim U[0, 2\pi]$, for $i = 1, \dots, m$. In our experiments, we compare the performance and computation times of both $\widehat{\text{MMD}}_k$ and $\widehat{\text{MMD}}_k^m$.

B A BRIEF SURVEY OF RELATED WORK

The literature about learning FCMs from data is vast. We recommend the books (Spirtes et al., 2000; Pearl, 2003; Peters et al., 2017) and surveys (Kalisch & Bühlmann, 2014; Heinze-Deml et al., 2017; Drton & Maathuis, 2017). FCM learning methods can be classified into bivariate and multivariate algorithms.

On the one, pairwise algorithms aim at orienting the cause-effect relation between two random variables (X, Y) by searching for asymmetries in the distribution $P(X, Y)$. The Additive Noise Model, or ANM (Hoyer et al., 2009), assumes an FCM with form $Y \leftarrow f(X) + E$, where the cause X is statistically independent from the noise E. Following these assumptions, the ANM performs one nonlinear regression in each direction, and prefers the one that produces residuals statistically independent from the alleged cause. The Post Non-Linear (PNL) model (Zhang & Hyvärinen, 2009) extends the ANM by allowing FCMs with form $Y \leftarrow g(f(X) + E)$, where g is a monotone function. The IGCI method (Daniusis et al., 2012) prefers the causal direction producing a cause distribution independent from the derivative of the causal mechanism. The LiNGAM method (Shimizu et al., 2006) leverages independent component analysis to orient linear cause-effect relations. The CURE

method (Sgouritsa et al., 2015) assumes that the distribution of the cause is independent from the causal mechanism, and prefers the causal direction that maximizes such independence. The GPI method (Stegle et al., 2010) fits a Bayesian model in both directions, and prefers as causal the one producing the largest marginal likelihood. The most similar work to this article is the one of Lopez-Paz & Oquab (2017), where a conditional generative adversarial network is trained to model a single equation from a FCM. Another set of methods, which include RCC (Lopez-Paz et al., 2015), NCC (Lopez-Paz et al., 2017), and Jarfo (Fonollosa, 2016) treat the cause-effect inference problem as a binary classification task. To this end, these methods need samples of causally-related variables, such as the ones pioneered by the competitions of Guyon (2013; 2014). However, these learning-based methods perform poorly when applied to cause-effect variables that differ from the training set. Statnikov et al. (2012); Mooij et al. (2016) reviews pairwise methods.

On the other hand, multivariate algorithms can be classified in to *constraint-based*, *score*, *hybrid*, and *pairwise* methods. First, constraint-based methods, such as the PC algorithm (Spirtes et al., 1999; Colombo et al., 2012) exploit conditional independences in data to construct an skeleton, and then orient v-structures. The constraint-based FCI algorithm (Spirtes et al., 2000) is an extension of the PC algorithm able to deal with hidden variables. All constraint-based methods perform conditional independence tests, which are usually implemented in terms of kernel methods (Zhang et al., 2012) and require an exponential amount of data in the number of conditioning variables to be reliable. Second, score methods search the space of causal graphs by minimizing a score function in a greedy way (for instance, by adding, removing, or reversing edges). Some examples of score methods are the GES (Chickering, 2002) and the Fast GES (Ramsey, 2015) algorithms. Third, hybrid methods (Drton & Maathuis, 2017; Nandy et al., 2015) combine constraint-based and score methods. Some examples include the MMHC (Tsamardinos et al., 2006) and GFCI (Ogarrio et al., 2016) algorithms.