The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

SELF: Structural Equational Likelihood Framework for Causal Discovery

Ruichu Cai,[1] Jie Qiao,[1] Zhenjie Zhang,[2] Zhifeng Hao[1, 3]

[1] School of Computer Science, Guangdong University of Technology, China
[2] Advanced Digital Sciences Center, Illinois at Singapore Pte. Ltd., Singapore
[3] School of Mathmatics and Big Data, Foshan University, China
cairuichu@gdut.edu.cn, qiaojie.chn@qq.com, zhenjie@adsc.com.sg, zfhao@gdut.edu.cn

Abstract

Causal discovery without intervention is well recognized as a challenging yet powerful data analysis tool, boosting the development of other scientific areas, such as biology, astronomy, and social science. The major technical difficulty behind the observation-based causal discovery is to effectively and efficiently identify causes and effects from correlated variables given the existence of significant noises. Previous studies mostly employ two very different methodologies under Bayesian network framework, namely global likelihood maximization and locally complexity analysis over marginal distributions. While these approaches are effective in their respective problem domains, in this paper, we show that they can be combined to formulate a new global optimization model with local statistical significance, called structural equational likelihood framework (or SELF in short). We provide thorough analysis on the soundness of the model under mild conditions and present efficient heuristic-based algorithms for scalable model training. Empirical evaluations using XGBoost validate the superiority of our proposal over state-of-the-art solutions, on both synthetic and real world causal structures.

Introduction

Causal discovery is well recognized as a challenging yet powerful data analysis tool (Pearl 2009; Spirtes, Glymour, and Scheines 2000), used to support a wide class of important applications, including biology (Grosse-Wentrup et al. 2016; Cai, Zhang, and Hao 2013a), computational astronomy (Schölkopf et al. 2016) and social science (Cai et al. 2017). Given a group of observation samples, causal discovery identifies the cause variables and effect variables, which explains the underlying mechanism of the physical world. While intervention is heavily exploited in the process of traditional causal discovery, recent research efforts mostly focus on analysis without intervention (Mooij et al. 2016b; Spirtes and Zhang 2016), due to the forbidden cost.

Causal diagram is commonly used to model the causal structure behind the multivariate observations, such that each variable is statistically determined by only a number of causal variables. The problem of causal discovery is therefore equivalent to the reconstruction of the causal structure, especially over the target effect variables. There are two general methodologies explored in the existing studies, which look into the problem in global and local views respectively. The approaches based on global view include the constraint-based methods (Spirtes, Glymour, and Scheines 2000; Pearl and Verma 1995), and the score-based approaches (Tsamardinos, Brown, and Aliferis 2006; Lam and Bacchus 1994; Ramsey et al. 2017). One common challenge behind the global view methods is the problem of Markov equivalence class (Andersson et al. 1997), such that certain causal graphical structures are indistinguishable based on marginal distribution information only, even when there is no noise injected into the stochastic generative process.

The approaches based on local view attempt to tackle the problem by looking into the coding complexity of the variables in the generative model, taking errors and noises into account. Specifically, the complexity of the generative process from cause variables to effect variables is supposed to be lower than the synthetic generative process on reversed direction. While the general Kolmogorov complexity is not computable (Janzing and Schölkopf 2010), a number of simpler complexity models are proposed in the literature, based on various assumptions over the generative process beneath the distributions, post-nonlinear model (Zhang and Hyvärinen 2009), additive noise model (Hoyer et al. 2009; Peters et al. 2014) and Information-geometric approach (Janzing et al. 2012). Most of these approaches look at a small number of variables, two in most cases, because of limited scalability rooted at the theory and algorithms (Mooij et al. 2016a). While a variety of scalability enhance schemes are proposed in the literature (Xie and Geng 2008; Cai, Zhang, and Hao 2013b), the improvement is limited, especially when local views over the variables include inaccurate causal results.

We believe these two general methodologies based on global and local views are not contradicted but complement to each other. Local views provide more accurate insight into the causal directions between individual pairs of variables, while global views are capable of correcting minor mistakes of local views by choosing most reasonable network structures over various options of directed edges extracted from the local views. The combination of these methodologies, however, is non-trivial. The key is to design an optimization framework with appropriate objective and constraints, unifying the local and global views of the Bayesian network in a consistent way.

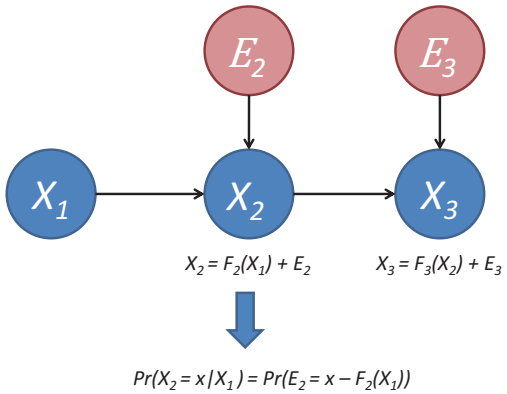


Figure 1: The causal graphical model consists of three variables $\{X_1, X_2, X_3\}$. The underlying generative process runs by injecting noises into the variables after a deterministic function, e.g., $F_2(\cdot)$ for X_2, is applied based on the graphical model. Therefore, the probability of an observation on a variable is equivalent to the probability of an observation on the noise variable.

In Figure 1, we illustrate the general motivation behind our new framework. In our example, the samples are generated following the graphical model over 3 variables, with noises independently injected into the variables. Given a group of structural equations corresponding to the underlying generative processes of the variables, the distribution of the observations on the variables is fully determined by the distribution of the noises. The likelihood of the observations is thus maximized when the noise estimations, based on optimized structural equations, over the variables provide the best match to the expected distribution of the noises. On the other hand, the random noises are supposed to be statistically independent of the cause variables, following the local views on the generative process. Therefore, we propose a novel structural equational likelihood framework (or SELF in short), which focuses on the noise estimation, by maximizing the global likelihood of the entire Bayesian network while preserving local statistical independence between noise and cause variables.

Based on the SELF framework, this paper covers a suite of technical contributions, including: (1) a mathematical formalization of the structural equational likelihood framework; (2) an effective and efficient implementation based on XGBoost to support a wide class of candidate regression models for the structural equations; (3) a theoretical analysis on the soundness of the model under mild conditions on the Bayesian network; (4) extensive empirical evaluations over linear non-Gaussain and nonlinear additive noise models by using both synthetic and real world causal structures.

Structural Equational Likelihood Framework

This section introduces the mathematical formalization of our structural equational likelihood framework (SELF).

We use upper case letters to denote variables and lower case letters to denote concrete values. G is the ground-truth causal graph, with vertices representing variables, i.e., $X = \{X_1, X_2, ..., X_n\}$, and directed edges representing causal directions, i.e., $\{X_i \to X_j\}$. A variable X_i is called a parent of X_j, if $X_i \to X_j$ holds. Each variable X_i corresponds to a distribution $\Pr(X_i = x)$ indicating the probability of $X_i = x$ for any valid x. The conditional distribution $\Pr(X_i \mid P_i)$ indicates the probability of observations on X_i with conditions on the values of all its parents. Give the causal graph G and the validity of the causal Markov condition(Spirtes, Glymour, and Scheines 2000; Pearl 2009), the joint distribution $\Pr(X)$ can be decomposed as the product of conditional distributions, as $\Pr(X) = \prod_{i=1}^{n} \Pr(X_i \mid X_{P_i})$, where X_{P_i} includes all parents of X_i in G. Following the common practice in causality research, we simply assume the causal graph is faithful to the result distribution. Given a group of observations $O = \{o_1, o_2, \ldots, o_m\}$, with each o_j as a n-dimensional vector $(o_{j,1}, \ldots, o_{j,n})$, we use o_{j,P_i} to denote the sub-vector of o_j containing values on variables in X_{P_i} only. Combined with the joint distribution $\Pr(X)$ and the causal graph G, the log-likelihood of the observations is calculated as

$$\mathcal{L}(G; O) = \sum_{j=1}^{m} \sum_{i=1}^{n} \log\left(\Pr(X_i = o_{j,i} \mid X_{P_i} = o_{j,P_i})\right) \quad (1)$$

Given the definition of log-likelihood over the observations, it is straightforward to design algorithms searching for optimal Bayesian network structure maximizing the likelihood. However, existing studies imply that such optimization may not return true causality structures, because of the existence of graphical structures rending exactly the same likelihood. These graphical structures are called Markov equivalence classes. Recall our example in Figure 1. Without the effect of noise, i.e., zero noise in the generative process, the log-likelihood of the ground-truth structure is identical to other structures, e.g., $X_1 \leftarrow X_2 \to X_3$.

To address the challenge, we introduce the idea of structural equation as well as probabilistic noise into the likelihood maximization scheme, in order to dissolve the ambiguity from the Markov equivalent classes. We use the additive noise model (Hoyer et al. 2009) $X_i = F_i(X_{P_i}) + E_i$ to present the causal mechanism behind the data, where F_i is the causal function of X_i belonging to a wide class of functions. Particularly, the randomized noise variable E_i is independent of the causal variables in X_{P_i}, i.e., $E_i \perp\!\!\!\perp X_{P_i}$.

Given the presence of the structural equation F_i for variable X_i and the assumption of independence between E_i and the parents of X_i, it is easy to verify the correspondence between the probability of an observation o_j and the probability of the noise observation.

$$\begin{aligned} &\Pr\left(X_i = o_{j,i} \mid X_{P_i} = o_{j,P_i}\right) \\ &\overset{X_i = F_i(X_{P_i}) + E_i}{=\!=\!=} \Pr\left(E_i = o_{j,i} - F_i(o_{j,P_i}) | X_{P_i}\right)\right) \quad (2) \\ &\overset{X_{P_i} \perp\!\!\!\perp E_i}{=\!=\!=} \Pr\left(E_i = o_{j,i} - F_i(o_{j,P_i})\right) \end{aligned}$$

Let $S = \langle G, F \rangle$ denote the causal structure and its corresponding structural equation, the log-likelihood over the

observations given in Formula 1 could be converted into the log-likelihood over the noise estimations, such that the maximization target is the noises instead of the observations, as

$$\mathcal{L}(S;O) = \sum_{j=1}^{m}\sum_{i=1}^{n} \log(\Pr(E_i = o_{j,i} - F_i(o_{j,P_i})) \quad (3)$$

The major benefit of the new log-likelihood formulation is the binding of likelihood with the noises. It further enables us to associate the noises with the independence constraints from the additive noise assumption, in the sense that these constraints can be directly achieved in the optimization formulation by manipulating the noise estimations. Consequently, it is straightforward to design a new optimization framework, as formalized in the following definition of structural equational likelihood framework (SELF).

Definition 1. *Given the observations O, construct a graphical model G and corresponding structural equations $\{F_i\}$ for all variables in X, to maximize the log-likelihood of the observations in Eq. 3, under the assumption of independence between the noise E_i and P_i for each variable X_i.*

In the rest of this section, we focus on the implementation of causal discovery algorithm under SELF. On a data set with limited sample size, the proposed $\mathcal{L}(S;O)$ tend to produce excessive redundant causal edges, when the optimization does not include any regularization on the complexity of the causal structure. To mitigate this effect, we introduce the Bayesian Information Criterion (BIC) penalty $\frac{d_i \log(m)}{2}$ into the $\mathcal{L}(S;O)$, where d_i is the number of coefficients used in the estimation of X_i. The new objective with BIC penalty is given in Formula 4.

$$\mathcal{L}_{\mathcal{B}}(\mathcal{S};\mathcal{O}) = \sum_{i=1}^{n}\left(\sum_{j=1}^{m} \log(\Pr(E_i = o_{j,i} - F_i(o_{j,P_i}))) - \frac{d_i \log(m)}{2}\right) \quad (4)$$

The maximization of the above objective function can be solved by an augmented optimization algorithm with two steps in each iteration, i.e., $\max \mathcal{L}_{\mathcal{B}}(S;O) = \max_G \sup_F \mathcal{L}_{\mathcal{B}}(\langle G, F\rangle; O)$. The first step is the estimation of $sup_F \mathcal{L}_{\mathcal{B}}(\langle G, F\rangle; O)$. The second step is the searching of the best causal graph with highest $\max_G \mathcal{L}_{\mathcal{B}}(\langle G, F\rangle; O)$.

The optimization of $sup_F \mathcal{L}_{\mathcal{B}}(\langle G, F\rangle; O)$ is solved by adopting the following two-step procedure. First, a regression with L_2 norm of the residual (i.e., $\sum_{i=1}^{n} ||E_i||_2$) is conducted to obtain the estimated noise E_i. Second, the kernel density estimation is employed to approximate the distribution of the noise. Here we do not directly optimize the entropy, because of the following two reasons: 1) there is no regression method whose objective function is the entropy of the residual, 2) the minimization of the entropy is equivalent to the minimization of the variance in a variety of distributions, such as the exponential family (Ahmed and Gokhale 1989). Considering the various formals of F in the real world applications, we employ XGBoost in this work as the function class for candidate regression models.

Algorithm 1 Hill-Climbing Based Causal Structure Search

Input: Observation O
Output: Causal structure $\langle G, F\rangle$
1: $G \leftarrow empty\ graph$, $F \leftarrow null\ function$
2: Initialize $\mathcal{L}$ according to Formula 4, $\mathcal{L}^* \leftarrow 0$
3: **while** $\mathcal{L}^* \leq \mathcal{L}$ **do**
4: **for** every $G' \in \mathcal{V}(G)$ **do**
5: Updating F_i, $\mathcal{L}_i'$ for $X_i \in \Delta(G, G')$
6: $\mathcal{L}' \leftarrow \sum_i \mathcal{L}_i'$
7: **end for**
8: $\langle G^*, F^*, \mathcal{L}^*\rangle \leftarrow \langle G', F', \mathcal{L}'\rangle$ with largest $\mathcal{L}'$
9: **if** $\mathcal{L}^* > \mathcal{L}$ **then**
10: $\langle G, F, \mathcal{L}\rangle \leftarrow \langle G^*, F^*, \mathcal{L}^*\rangle$
11: **end if**
12: **end while**
13: **return** $\langle G, F\rangle$

In the searching of G with highest $\max_G \mathcal{L}_{\mathcal{B}}(\langle G, F\rangle; O)$, the hill-climbing based local search algorithm is used. In the hill-climbing algorithm, each iteration searches around the vicinity of the current causal graph structure G with only one causal edge added, deleted or reversed, denoted by $\mathcal{V}(G)$. Because the $\mathcal{L}_{\mathcal{B}}(S;O)$ can be decomposed to the sum of the likelihood of the variables, the objective function can be efficiently estimated by conducting local updating scheme on the nodes belongs to the incremental set $\Delta(G, G')$ between G and G'. The variable set $\Delta(G, G')$ contains all the variables whose parents are different in G and G'. The local updating rule for X_i is as follows: $\mathcal{L}_{\mathcal{B}_i}(S;O) = \sum_{j=1}^{m} \log(\Pr(E_i = o_{j,i} - F_i(o_{j,P_i}))) - \frac{d_i \log(m)}{2}$.

The details of the hill-climbing algorithm are provided in Algorithm 1. Specifically, the algorithm first initializes $\mathcal{L}$ with the empty graph G and null function F (Line 1-2). Then search the best fit graph iteratively by performing the $add, delete, reverse$ operations on G to generate the candidate graphs in $\mathcal{V}(G)$ at each iteration (Line 4). In each iteration, in order to calculate the score for each candidate G', the algorithm locally updates the nodes belongs to the set $\Delta(G, G')$ (Line 5). The algorithm then identifies the highest score and its corresponding causal structure among the candidate graphs (Line 8). It repeats the local search process until the score is no longer improved.

Soundness of SELF

In this section, we prove that correct causal structure always renders highest score in SELF. Secondly, we prove the structure with the highest SELF score is the correct causal structure of the data, under certain mild and reasonable assumptions. Note that the conclusions are valid for both $\mathcal{L}(S;O)$ and $\mathcal{L}_{\mathcal{B}}(S;O)$ as the objective function, because the BIC penalty of $\mathcal{L}_{\mathcal{B}}(S;O)$ includes $\frac{d_i log(m)}{2m}$ for each sample and it generally converges to 0 when there are sufficient samples.

Theorem 1. *Given a large enough observation O generated from S, $\mathcal{L}(S;O) \geq \mathcal{L}(S';O)$ holds for any S'.*

Proof. The proof is given in Equation 5. The first three equal-

ities are based on Equation 2 and 3. The fourth equality is based on the sufficiently large observations. The fifth equality is based on the definition of KL-divergence. The last inequality is based on the property of KL-divergence.

$$
\begin{aligned}
&\mathcal{L}(S;O) - \mathcal{L}(S';O) \\
&= \sum_{j=1}^{m}\sum_{i=1}^{n} \log\left(\frac{\Pr(E_i = o_{j,i} - F_i(o_{j,P_i}))}{\Pr(E_i = o_{j,i} - F_i'(o_{j,P_i'}))}\right) \\
&= \sum_{j=1}^{m}\sum_{i=1}^{n} \log\left(\frac{\Pr(X_i = o_{j,i} | P_i = o_{j,P_i})}{\Pr(E_i = o_{j,i} - F_i'(o_{j,P_i'})}\right) \\
&= m\sum_{j=1}^{m} \log\left(\frac{\Pr(X = o_j)}{\prod_{i=1}^{n}\Pr(E_i = o_{j,i} - F_i'(o_{j,P_i'}))}\right) \qquad (5) \\
&= mE_{o\sim S}\log\left(\frac{\Pr(X = o)}{\prod_{i=1}^{n}\Pr(E_i = o_i - F_i'(o_{P_i'})}\right) \\
&= mKL\left(\Pr(X = o) || \prod_{i=1}^{n}\Pr(E_i = o_i - F_i'(o_{P_i'})\right) \\
&\geq 0
\end{aligned}
$$

□

Intuitively, Theorem 1 shows that $\mathcal{L}(S;O)$ achieves the highest likelihood, because the $\Pr(E_i = o_{j,i} - F_i(o_{j,P_i}))$ reflects the true distribution of the data. However, Theorem 1 does not ensure the structure S with highest $\mathcal{L}(S;O)$ is the correct causal structure.

In the following, to facilitate the further analysis, we prove the equivalence between the SELF score and the minimization of the entropy of noise over the variables in Lemma 1. With the help of Lemma 1, Lemma 2 and Theorem 2 proves the final conclusion that the structure with the highest SELF score is the correct causal structure of the data.

Lemma 1. *Assume that the samples are independent and identically distributed,* $\arg\max_S \mathcal{L}(S;O) = \arg\min_S \sum_{i=1}^{n} H(E_i|S)$ *holds.*

Proof. We have $\mathcal{L}(S;O) = m\sum_{i=1}^{n}\frac{1}{m}\sum_{j=1}^{m}\log\left(\Pr\left(E_i = o_{j,i} - F_i(o_{j,P_i})\right)\right) = m\sum_{i=1}^{n} E(\log(\Pr(E_i))) = -m\sum_{i=1}^{n} H(E_i|S)$. The second equality holds based on the assumption that the sample size is large enough, and the last equality is based on the definition of entropy. □

This lemma shows that the maximization of $\mathcal{L}(S;O)$ is equivalent to the minimization of the entropy of noise. In the following, we employ this property to investigate the relationship between $\mathcal{L}(S;O)$ and the casual structure with the help of the assumption 1.

Assumption 1. *For* $\forall X_k \in X_{P_i}$ *and* $\forall X_z \notin X_{P_j}$, $H(E_j|X_{P_j}) - H(E_j|X_{P_j\cup\{z\}}) \ll H(E_i|X_{P_i-\{k\}}) - H(E_i|X_{P_i})$ *holds.*

The Assumption 1 concerns the relative gap of the conditional entropy when parent and non-parent variables are included in condition set. This assumption ensures the correct parent node reduces more on entropy than the non-parental nodes do. The rationality of this assumption is based on the following two observations: 1) $H(E_j|X_{P_j}) - H(E_j|X_{P_j\cup\{z\}}) \leq H(E_j|X_{P_j})$ holds. When given the correct parents set X_{P_j} of X_j, $H(E_j|X_{P_j})$ is a small variable, thus any non-parent variable has ignorable effect on reducing entropy of X_j. 2) Any variable $X_k \in X_{P_i}$ can greatly reduce the entropy of X_i, because the state of X_i is determined by its parents. Thus, the above assumption is reasonable. We also tested the applicability of Assumption 1 in the experiment section.

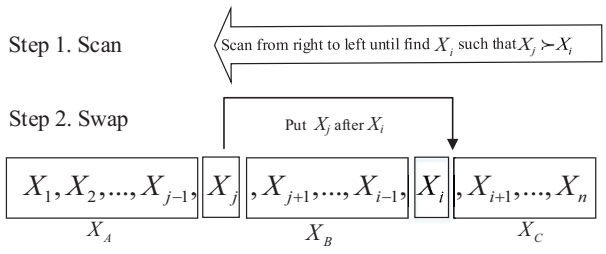


Figure 2: Sketch diagram of the local swap

Let $T(G)$ denote the topological order of the directed acyclic graph G. $T(G)$ is compatible with the ground-truth causal structure if and only if $X_i \preceq X_j$ holds for each $X_i \to X_j$ in the ground-truth causal structure.

Lemma 2. *When* $\sum_i H(E_i|S)$ *reaches its minimum on* $S = \langle G, F\rangle$, $T(G)$ *is compatible with the ground-truth causal structure.*

Proof. (Proof by Contradiction.) Assume $\sum_i H(E_i|S)$ reaches its minimum on $\langle G, F\rangle$, but $T(G)$ is not compatible with the ground-truth causal structure. We will prove that there exist a series of local swaps on $T(G)$ which satisfy: 1) $T(G)$ after swapping is compatible with ground-truth, and 2) $\sum_i H(E_i|S)$ decreases after each local swap. The existing of the above local swaps implies there exists a new causal structure G' whose $\sum_i H(E_i|S)$ is smaller than that of G, which is contradicted with the $\sum_i H(E_i|S)$ reaches its minimum on $\langle G, F\rangle$. This finishes the proof. The details of the local swaps are given as follows.

Without loss of generality, let $T(G)$ be $X_1 \preceq X_2, \preceq ... \preceq X_n$. Assume the $T(G)$ is not compatible with ground-truth, we will use the following two steps local swap method to transform $T(G)$ to the correct order, as shown in Figure 2. 1) Scan from the end of $T(G)$ until find the pair of variable satisfies $X_i \preceq X_j$ and $i > j$, which means that X_j is falsely placed before X_i. 2) Put X_j after X_i. This two-step local swap is repeated until all the variables are correctly placed.

Next we will prove that $\sum_i H(E_i|S)$ decreases when put X_j after X_i. As shown in the figure 2, $T(G)$ is divided into the following five parts $\{X_A\}, \{X_j\}, \{X_B\}, \{X_i\}, \{X_C\}$. The gap of the score after one swap operation is given in Equation 6. For the simplicity of the notation, we consider $H(X_B|X_A)$ for every nodes in X_B, the latter nodes

are conditioned by the former nodes. The mutual information is equal to zero when given the condition set is a super set of the parent set. Then the most of the mutual information equal to zero in equation 6. Moreover, because the X_i is the largest index such that $X_i \preceq X_j$ where $i > j$ and the nodes of X_B are recursively sorted, each parents of $\{\{X_j\}, \{X_B\}, \{X_i\}, \{X_C\}\}$ are all included in its precedent nodes. The last inequality is based on the Assumption 1, because $H(E_i|X_{A,B}) - H(E_i|X_{A,B,j})$ is much less than both $H(E_j|X_A) - H(E_j|X_{A,B,i})$ and $H(E_B|X_{A,j}) - H(E_B|X_A)$.

$$\begin{aligned}
&\big(H(E_j|X_{A,B,i}) + H(E_B|X_A) + H(E_i|X_{A,B})\big) \\
&- \big(H(E_j|X_A) + H(E_B|X_{A,j}) + H(E_i|X_{A,j,B})\big) \\
=&H(E_j|X_{A,B,i}) - H(E_j|X_A) - I(E_j, X_A) \\
&+ H(E_B|X_A) - H(E_B|X_{A,j}) \\
&+ H(E_i|X_{A,B}) - H(E_i|X_{A,j,B}) \\
<&0
\end{aligned} \tag{6}$$

□

Theorem 2. *When $\sum_i H(E_i|S)$ reaches its minimum on $S = \langle G, F\rangle$, $\langle G, F\rangle$ is the correct causal structure.*

Proof. (Proof by Contradiction.) Based on lemma 2, $T(G)$ is compatible with the ground-truth when $\sum_i H(E_i|S)$ reaches its minimum on $\langle G, F\rangle$. Without loss of generality, let $T(G)$ be $X_1 \preceq X_2, \preceq \ldots \preceq X_n$. Suppose there exists a $\langle P_i', F_i'\rangle$ where F is non-degenerate function and satisfies: 1) $\langle P_i', F_i'\rangle \neq \langle P_i, F_i\rangle$, 2) $\sum_i H(E_i|S') < \sum_i H(E_i|S)$.

According to $\langle P_i', F_i'\rangle \neq \langle P_i, F_i\rangle$ and structural equation model assumptions, we have the first inequality, and the first equality is based on the definition of E_i, and the second equality is because that $F(P_i)$ is determined function of X_{P_i}.

$$\sum_i H(E_i|S') \geq \sum_i H(E_i|X_{P_i'}) = \sum_i H(X_i|X_{P_i'}) \tag{7}$$

Based on the causal Markov assumption, we have

$$\sum_i H(X_i|X_{P_i'}) \geq \sum_i H(X_i|X_{P_1,P_2,\cdots,P_{i-1}}) = \sum_i H(X_i|X_{P_i}) \tag{8}$$

Based on structural equation model assumption, $X_i = F(X_{P_i}) + E_i$ and $E_i \perp\!\!\!\perp X_{P_i}$, we have

$$\sum_i H(X_i|X_{P_i}) = \sum_i H(E_i|S) \tag{9}$$

Combining Inequality 7, Inequality 8 and Inequality 9, we have $\sum_i H(E_i|S') \geq \sum_i H(E_i|S)$, which contradict with the supposition $\sum_i H(E_i|S') < \sum_i H(E_i|S)$. This finishes the proof. □

Experiments and Discussions

Experiment Settings

To investigate the effectiveness and of genericity of SELF, the algorithms are tested on both linear non-Gaussian and nonlinear data, generated from synthetic and real world causal structures. Specifically, the linear non-Gaussian data are generated according to the following linear structural equations, $x_i = \sum_{j \in P_i} w_j x_j + e_i$ with random coefficients $w \sim U(0.5, 1) \cup U(-0.5, -1)$, and noise $e_i \sim sub - Gaussian$; the nonlinear data are generated according to the following nonlinear function, $y = scale(a_1 b_1 x^2 + a_2 b_2 x^3 + a_3 b_3 x^4 + a_4 b_4 sin(x) + a_5 b_5 sin(x^2)) + e_i$, with $b_i \in \{0, 1\}$ is a random indicator, $a_i \sim U(-3, 3)$ is the random weight of each component, $e_i \sim sub - Gaussian$ is the noise, and $scale$ is a normalized function. These data generation processes follow the practice in state-of-the-art research works, e.g., (Shimizu et al. 2006) and ANM (Hoyer et al. 2009).

In linear data, we compare SELF to three state-of-the-art algorithms, include LiNGAM (Shimizu et al. 2006), DLiGANM(Shimizu et al. 2011), and HCBN (a hill-climbing based Bayesian network search algorithm for linear data) (Scutari 2009). We reuse the parameter settings in LiNGAM and DLiNGAM based on the descriptions from their original papers, and the implementation and parameter settings of HCBN are based on bnlearn package in R. For SELF, XGBoost-GBLinear (Chen and Guestrin 2016) is employed as the linear regression method and kernel density estimation (Parzen 1962) is used to estimate the probability density function of the noise.

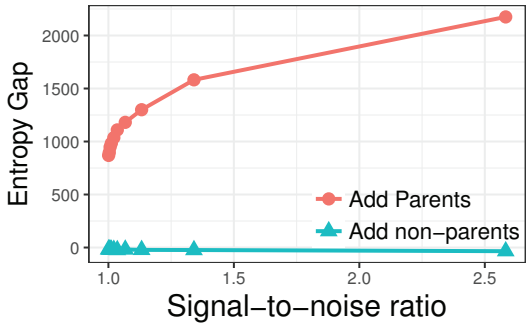
In nonlinear data, we only compare our work with MMPC-ANM, because there is no other methods applicable to nonlinear data with multiple variables. The implementation of MMPC-ANM is based on bnlearn package and CompareCausalNetworks package(Heinze-Deml and Meinshausen 2016). For SELF, XGBoost-GBTree (Chen and Guestrin 2016) is employed as the regression method. The estimation of the probability density function of noise is identical to the linear case.

In all the following experiments, recall, precision, and F1 are recorded for all the algorithms as the evaluation metrics. All the reported results are based on at least 10 runs of the respective algorithms. The implementation of SELF can be found in CRAN [1].

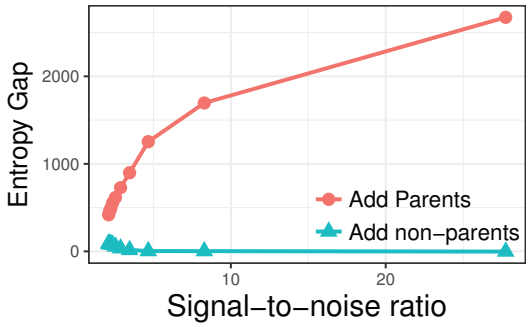
Synthetic Structures

In this part, we design a series of controlled experiments on the random causal structures with given samples size, variable size, and average indegree. The ranges of the above three parameters are as follows: the number of variables=$\{20, 30, \mathbf{40}, 50, 60\}$, the number of samples=$\{1000, 2000, \mathbf{4000}, 6000, 8000\}$, and the average indegree=$\{0.5, 1, \mathbf{1.5}, 2, 2.5\}$. The default setting of the parameters is marked in bold. Note that the above parameter setting reflects the real world causal structures, for example, the range of average indegree covers all the real world structures given in Table 1.

[1] https://cran.r-project.org/web/packages/SELF/index.html

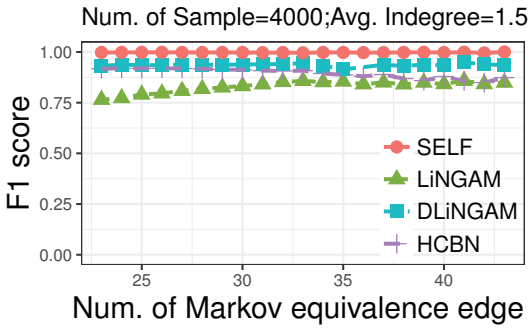


(a) Linear Data.

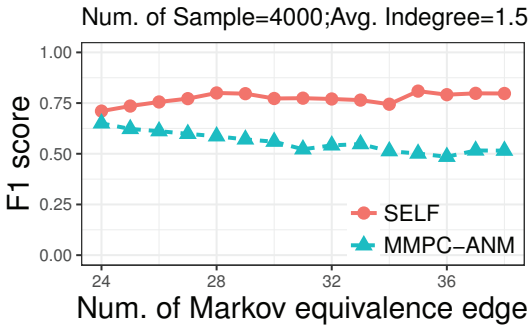


(b) Nonlinear Data.

Figure 3: The universality of Assumption 1.



(a) Linear Data.



(b) Nonlinear Data.

Figure 4: Results on the Different Markov Equivalents Edge.

Universality of Assumption 1: Firstly, we design a set of controlled experiments to verify the universality of Assumption 1, by comparing two gaps given in Assumption 1 on a set of data with different signal-noise ratio. The signal-noise ratio is controlled by the relative weight of noise. Fig. 3 shows that adding a parent node is always better than adding a non-parent node to a correct parent set, even when the signal-to-noise ratios are below 1. This verifies the applicability of the assumption.

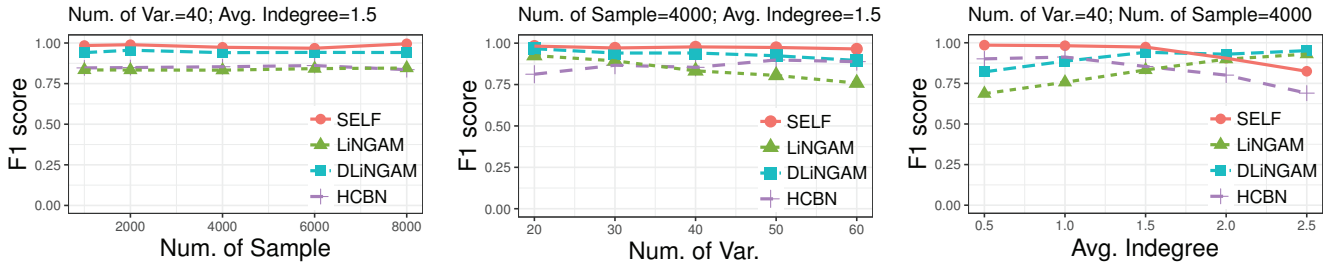
Effects of Markov equivalence class: Fig. 4 shows the effect of Markov equivalence class on the different methods. The algorithms are tested on random structures with various number of Markov equivalence edges. Here, Markov equivalence edges refer to the edges, whose direction can not be determined by V-Structures. In the linear experiment 4(a), the performance of the three structural equation model based methods (i.e., SELF, LiNGAM, and DLiNGAM) are robust to the number of Markov equivalence edges, which the F1 score of traditional likelihood based method (HCBN) decreases rapidly. This is because the traditional likelihood based method can not distinguish the Markov equivalence class. As a structural equational likelihood framework, SELF benefits from both the advantages of the structural equational model and Bayesian score based method. Similar results are observed in nonlinear experiments in Figure4(b).

Sensitivity to Structures: Figure 5 shows the results on the linear data. Generally, SELF outperforms the other three state-of-the-art methods on all the settings. In detail, Figure 5(a) shows the performance of the methods under different number of samples, the results reflect that 1,000 samples are enough for all the methods, and SELF outperforms the other three methods across all the sample size configurations. Figure 5(b) compares the methods under different number of variables, the improvement gap between SELF and LiNGAM/DLiNGAM grows with the number of variables, which validates the scalability of SELF to multiple variables. Figure 5(c) compares the methods under different average indegree. An interesting observation is that SELF and HCBN work well on the sparse causal structures, while LiNGAM and DLiNGAM prefer denser structures. This turns out to be an advantage of SELF, because real world causal structures are known to be sparse (Pearl 2009).

Fig. 6 shows the results on the nonlinear data. Similar to the results on linear data, SELF significantly outperforms the baseline methods on all configurations. A different phenomenon between the linear and nonlinear data is shown in Figure 6(a). On the nonlinear data, SELF needs more samples to obtain robust results, because of the nonlinear regression employed in SELF. The increase of the gap between SELF and MMPC-ANM with the sample size also reflects the ability of SELF on the exploration of higher order information beneath the data.

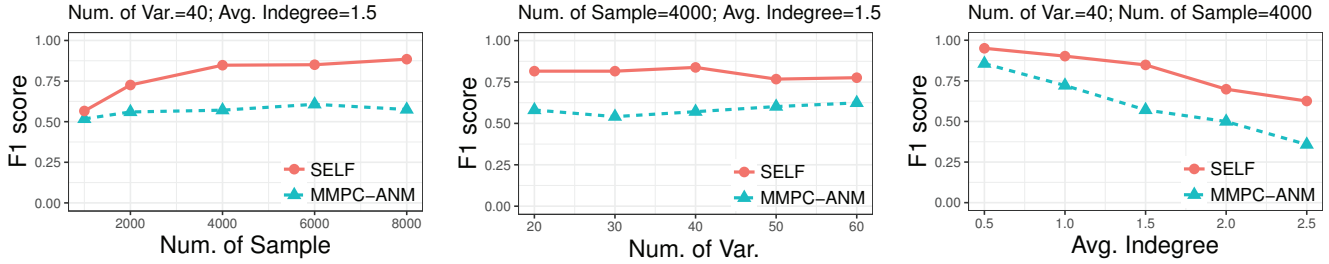
Real world Structures

In this part, we explore the performance of the algorithms on four frequently used real world structures (Scutari 2009). The statistics of the structures are given in Table 2. The other settings, i.e., the sample size, data generation function,



(a) Sensitivity to the number of samples. (b) Sensitivity to the number of variables. (c) Sensitivity to the average indegree.

Figure 5: Sensitivity Analysis on the Linear Data.



(a) Sensitivity to the number of samples. (b) Sensitivity to the number of variables. (c) Sensitivity to the average indegree.

Figure 6: Sensitivity Analysis on the Nonlinear Data.

Table 1: Statistics of the Real world Structures.

Structure	nodes	Edges	Avg deg.	Max deg.
Child	20	25	0.63	2
Alarm	37	46	1.25	4
Win95pts	76	112	1.48	7
Pathfinder	135	200	1.48	5

are identical to those in the experiments on the synthetic structures.

The results on real world structures also verify the effectiveness of SELF. As shown in Table 2 and Table 3, SELF outperforms the baseline methods on both linear and nonlinear data, over all the structures. In detail, the advantage of the SELF is more significant over problem domains with larger scale. It proves the excellent scalablity of SELF. Another important conclusion on linear data is that the improvement of the precision is much higher than that of recall. This reflects the capability of SELF to distinguish the correct causal structure from the Markov equivalence class.

Conclusion

In this work, we present SELF, a structural equational likelihood framework, together with a hill climbing based causal structure discovery algorithm, and discussions on the soundness of SELF in theory. Our experimental results validate the effectiveness and genericity of the proposed framework and algorithms. By employing the likelihood function globally and estimating the structural equation model locally, SELF provides a unified and theoretically robust methodology for causal structure exploration. The success of SELF also verifies that the global approaches and the local approaches are complementary to each other. Future work includes extending SELF to other causal mechanisms compatible with structural equation models, and accelerating the hill climbing search by parallel evaluation of the objective function.

Acknowledgments

This research is supported in part by NSFC-Guangdong Joint Found (U1501254), Natural Science Foundation of China (61472089), Natural Science Foundation of Guangdong (2014A030306004, 2014A030308008), Science and Technology Planning Project of Guangdong (2015B010108006, 2015B010131015), Guangdong High-level Personnel of Special Support Program (2015TQ01X140) , Pearl River S&T Nova Program of Guangzhou (201610010101), and Science and Technology Planning Project of Guangzhou (201604016075). This research is also supported by the National Research Foundation, Prime Ministers Office, Singapore, under its Campus for Research Excellence and Technological Enterprise (CREATE) programme.

References

Ahmed, N. A., and Gokhale, D. 1989. Entropy expressions and their estimators for multivariate distributions. *IEEE Transactions on Information Theory* 35(3):688–692.

Andersson, S. A.; Madigan, D.; Perlman, M. D.; et al. 1997. A characterization of markov equivalence classes for acyclic digraphs. *The Annals of Statistics* 25(2):505–541.

Table 2: Results on real world structure with linear data.

Dataset	F1				Recall				Precision			
	SELF	LiNGAM	DLiNGAM	HCBN	SELF	LiNGAM	DLiNGAM	HCBN	SELF	LiNGAM	DLiNGAM	HCBN
Child	**0.98**	**0.98**	0.95	0.58	**1.00**	**1.00**	**1.00**	0.65	**0.96**	0.95	0.92	0.52
Alarm	**0.98**	0.43	0.94	0.52	0.99	0.76	**1.00**	0.64	**0.96**	0.31	0.88	0.44
Win95pts	**0.95**	0.56	0.88	0.80	0.97	0.88	**1.00**	0.91	**0.93**	0.42	0.79	0.71
Pathfinder	**0.91**	0.86	0.85	0.73	0.95	**0.96**	**0.96**	0.83	**0.87**	0.77	0.76	0.64

Table 3: Results on real world structure with nonlinear data.

Data	F1		Recall		Precision	
	SELF	ANM	SELF	ANM	SELF	ANM
Child	**0.71**	0.26	**0.60**	0.40	**0.88**	0.19
Alarm	**0.79**	0.53	**0.74**	0.59	**0.85**	0.48
Win95pts	**0.77**	0.47	**0.71**	0.49	**0.86**	0.45
Pathfinder	**0.88**	0.15	**0.90**	0.08	0.86	**1.00**

Cai, R.; Zhang, Z.; Hao, Z.; and Winslett, M. 2017. Understanding social causalities behind human action sequences. *IEEE Trans. Neural Netw. Learning Syst.* 28(8):1801–1813.

Cai, R.; Zhang, Z.; and Hao, Z. 2013a. Causal gene identification using combinatorial v-structure search. *Neural Networks* 43:63–71.

Cai, R.; Zhang, Z.; and Hao, Z. 2013b. SADA: A general framework to support robust causation discovery. In *ICML*, 208–216.

Chen, T., and Guestrin, C. 2016. Xgboost: A scalable tree boosting system. In *SIGKDD*, 785–794.

Grosse-Wentrup, M.; Janzing, D.; Siegel, M.; and Schölkopf, B. 2016. Identification of causal relations in neuroimaging data with latent confounders: An instrumental variable approach. *NeuroImage* 125:825–833.

Heinze-Deml, C., and Meinshausen, N. 2016. *CompareCausalNetworks: Interface to Diverse Estimation Methods of Causal Networks*. R package version 0.1.5.

Hoyer, P. O.; Janzing, D.; Mooij, J. M.; Peters, J.; and Schölkopf, B. 2009. Nonlinear causal discovery with additive noise models. In *NIPS*, 689–696.

Janzing, D., and Schölkopf, B. 2010. Causal inference using the algorithmic markov condition. *IEEE Trans. Information Theory* 56(10):5168–5194.

Janzing, D.; Mooij, J.; Zhang, K.; Lemeire, J.; Zscheischler, J.; Daniušis, P.; Steudel, B.; and Schölkopf, B. 2012. Information-geometric approach to inferring causal directions. *Artificial Intelligence* 182:1–31.

Lam, W., and Bacchus, F. 1994. Learning bayesian belief networks: An approach based on the mdl principle. *Computational intelligence* 10(3):269–293.

Mooij, J.; Peters, J.; Janzing, D.; Zscheischler, J.; and Schölkopf, B. 2016a. Distinguishing cause from effect using observational data: methods and benchmarks. *Journal of Machine Learning Research* 17(32):1–102.

Mooij, J. M.; Peters, J.; Janzing, D.; Zscheischler, J.; and Schölkopf, B. 2016b. Distinguishing cause from effect using observational data: Methods and benchmarks. *Journal of Machine Learning Research* 17:32:1–32:102.

Parzen, E. 1962. On estimation of a probability density function and mode. *The annals of mathematical statistics* 33(3):1065–1076.

Pearl, J., and Verma, T. S. 1995. A theory of inferred causation. *Studies in Logic and the Foundations of Mathematics* 134:789–811.

Pearl, J. 2009. *Causality: models, reasoning and inference*. Cambridge university press.

Peters, J.; Mooij, J. M.; Janzing, D.; Schölkopf, B.; et al. 2014. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research* 15(1):2009–2053.

Ramsey, J.; Glymour, M.; Sanchez-Romero, R.; and Glymour, C. 2017. A million variables and more: the fast greedy equivalence search algorithm for learning high-dimensional graphical causal models, with an application to functional magnetic resonance images. *International journal of data science and analytics* 3(2):121–129.

Schölkopf, B.; Hogg, D.; Wang, D.; Foreman-Mackey, D.; Janzing, D.; Simon-Gabriel, C.-J.; and Peters, J. 2016. Modeling confounding by half-sibling regression. *Proceedings of the National Academy of Science* 113(27):7391–7398.

Scutari, M. 2009. Learning bayesian networks with the bnlearn R package. *arXiv preprint arXiv:0908.3817*.

Shimizu, S.; Hoyer, P. O.; Hyvärinen, A.; and Kerminen, A. 2006. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research* 7(Oct):2003–2030.

Shimizu, S.; Inazumi, T.; Sogawa, Y.; Hyvärinen, A.; Kawahara, Y.; Washio, T.; Hoyer, P. O.; and Bollen, K. 2011. Directlingam: A direct method for learning a linear non-gaussian structural equation model. *Journal of Machine Learning Research* 12(Apr):1225–1248.

Spirtes, P., and Zhang, K. 2016. Causal discovery and inference: concepts and recent methodological advances. In *Applied Informatics*, volume 3, 1. Springer Berlin Heidelberg.

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

Tsamardinos, I.; Brown, L. E.; and Aliferis, C. F. 2006. The max-min hill-climbing bayesian network structure learning algorithm. *Machine learning* 65(1):31–78.

Xie, X., and Geng, Z. 2008. A recursive method for structural learning of directed acyclic graphs. *Journal of Machine Learning Research* 9:459–483.

Zhang, K., and Hyvärinen, A. 2009. On the identifiability of the post-nonlinear causal model. In *UAI*, 647–655.