SADA: A General Framework to Support Robust Causation Discovery

Ruichu Cai CAIRUICHU@GMAIL.COM
Faculty of Computer Science, Guangdong University of Technology, and State Key Laboratory for Novel Software Technology, Nanjing University, P.R. China

Zhenjie Zhang ZHENJIE@ADSC.COM.SG
Advanced Digital Sciences Center, Illinois at Singapore Pte. Ltd., Singapore

Zhifeng Hao ZFHAO@GDUT.EDU.CN
Faculty of Computer Science, Guangdong University of Technology, P.R. China

Abstract

Causality discovery without manipulation is considered a crucial problem to a variety of applications, such as genetic therapy. The state-of-the-art solutions, e.g. LiNGAM, return accurate results when the number of labeled samples is larger than the number of variables. These approaches are thus applicable only when large numbers of samples are available or the problem domain is sufficiently small. Motivated by the observations of the local sparsity properties on causal structures, we propose a general *Split-and-Merge* strategy, named SADA, to enhance the scalability of a wide class of causality discovery algorithms. SADA is able to accurately identify the causal variables, even when the sample size is significantly smaller than the number of variables. In SADA, the variables are partitioned into subsets, by finding *cuts* on the sparse probabilistic graphical model over the variables. By running mainstream causation discovery algorithms, e.g. LiNGAM, on the subproblems, complete causality can be reconstructed by combining all the partial results. SADA benefits from the recursive division technique, since each small subproblem generates more accurate result under the same number of samples. We theoretically prove that SADA always reduces the scale of problems without significant sacrifice on result accuracy, depending only on the local sparsity condition over the variables. Experiments on real-world datasets verify the improvements on scalability and accuracy by applying SADA on top of existing causation algorithms.

1. Introduction

Causality discovery plays an important role on a variety of scientific domains. Different from the mainstream statistical learning approaches, causality learning tries to understand the data generation procedure, rather than characterizing the joint distribution of the observed variables only. It turns out that understanding causality in such procedures is essential to predict the consequences of interventions, which is the key to a large number of applications, such as genetic therapy, advertising campaign design, etc.

From computational perspective, causation discovery is usually formulated with a graphical probabilistic model on the variables (Pearl, 2009), such that directed edges between variables indicate causation relationships. When it is unlikely to manipulate the samples in experiments, conditional independence testing is commonly employed to detect local causalities among the variables (Pearl, 2009; Spirtes et al., 2001). Despite of the successes of these approaches on small problem domains and large sample bases, they usually fail to find true causalities, when huge equivalent classes over the graphical probabilistic models render exactly the same conditional independence.

To tackle the difficulties in the problem of causal structure learning under non-experimental setting, re-

Proceedings of the 30th International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28. Copyright 2013 by the author(s).

searchers are recently resorting to asymmetrical relationship between the cause and effect variables under assumptions on the generation process. The discovery ability is dramatically improved, by exploiting linear assumption (Zscheischler et al., 2012), linear non-Gaussian assumption (Shimizu et al., 2006; 2011), nonlinear non-Gaussian assumption (Hoyer et al., 2008), discrete property (Peters et al., 2010), and so on. When the variables are correlated under linear relationships and the noises follow non-Gaussian distributions, for example, LiNGAM (Shimizu et al., 2006) and its variants (Shimizu et al., 2011) are known as the best causality inference algorithms. However, the scalability of LiNGAM and its variants is still questionable, since they heavily depend on the independent component analysis (ICA) during the computation. To return robust results from ICA, it is necessary to feed a large bulk of samples, which are expected to be no smaller than the number of variables.

Motivated by the common observations on the sparsity of causal structures, i.e. each variable usually only depends on a small number of parent variables, we derive a new general framework in this paper. The new framework helps the existing causation algorithms to get rid of difficulties on small sample cardinality in practice. Well designed conditional independence testings are conducted first, to partition the problem domain into small subproblems. With the same number of samples, existing causation algorithms could generate more robust and accurate results on these small subproblems. Partial results from all subproblems are finally merged together, to return a complete picture of causalities among all the variables. This framework is theoretically solid, as it always returns correct and complete result under the optimal setting. Our experiments on synthetic and real datasets verify the superior scalability and effectiveness of our proposal, when applied together with two mainstream causation analysis algorithms.

2. Related Work

Causality Bayesian network (CBN) is part of the theoretical background of this work. CBN is a special case of Bayesian network, whose edge direction presents the causality relations among the nodes (Pearl, 2009). CBN has been used to model the causal structure in many real-world applications, for example, the gene regulatory network (Friedman et al., 2000; Kim et al., 2004), causal feature selection (Cai et al., 2011).

Most of existing work try to explore the structure learning approach to learning the CBN, e.g. the well known PC algorithm (Spirtes et al., 2001; Kalisch & Bühlmann, 2007), Markov Blanket discovery methods (Zhu et al., 2007). These methods provide the skeleton of causal structures, i.e. parent-child pairs and Markov Blanket. However, these methods usually cannot distinguish causes from consequences, thus unable to output exact causalities.

Pearl is the pioneer of the causality analysis theory (Pearl, 2009). Since Pearl's Inductive Causality (Pearl & Verma, 1991), a large number of extensions are proposed. Most causality inference algorithms assume the acquisition of a sufficiently large sample base (Aliferis et al., 2010). Though there are studies aiming at the inference problem under small sample cardinality (Bromberg & Margaritis, 2009), the actual number of the samples used in their empirical evaluations remains significantly larger than the number of variables. Cai's study (Cai et al., 2013) is another attempt under this category to extend the method to the high dimensional gene expression data by exploring the local substructures. However, all these approaches, based on independence conditional testings, cannot distinguish two causality structures if they come from a so-called *Markov equivalence class* (Pearl, 2009), in which expensive intervention experiments were previously considered essential (He & Geng, 2008).

Recently, *Additive Noise Model* is proposed to break the limitation of the class of method purely under conditional independence testings, by exploiting the asymmetric property of the noises in the generative progress, which brings a gleam of dawn to resolve the causal equivalence problem. The Additive Noise Model highly depends on the type of noise and the form of causality. Existing studies on this line can be categorized based on the adopted assumption on the noise type and data generation mechanism. LiNGAM and its variants (Shimizu et al., 2006; 2011), for example, assume that the data generating process is linear and the noise distributions are non-Gaussian. The nonlinear non-Gaussian method (Hoyer et al., 2008) works when the data generating process is nonlinear. And discrete model (Peters et al., 2010) is proposed for the causal inference on domains with only discrete variables. There are other research studies on this topic, such as explaining the underlying theoretical foundation behind additive noise mode (Janzing et al., 2012), regression-based model inference (Mooij et al., 2009) and kernel independence test (Zhang et al., 2012). To the best of our knowledge, there is no existing work to address the problem of sample cardinality under Additive Noise Model.

3. SADA Framework

3.1. Preliminaries

Assume that all samples from the problem domain contain information on m different variable, i.e. $V = \{v_1, v_2 \ldots, v_m\}$. Let $D = \{x_1, x_2, \cdots, x_n\}$ denote an observation sample set. Each sample x_i is denoted by a vector $x_i = (x_{i1}, x_{i2}, \ldots, x_{im}, y_i)$, where x_{ij} indicates the value of the sample x_i on variable v_j and y_i is the target variable under investigation. If $\mathcal{P}$ is a distribution over the domain of variables in V, we assume that there exists a causal Bayesian network N faithful to the distribution $\mathcal{P}$. The network N includes a directed acyclic graph G, each edge in which indicates a conditional (in)dependent relationship between two variable nodes. Each edge is also associated with a conditional probability function which simulates conditional probability distribution of each variable given the values of the parent variables. Following the common assumption of existing studies, we only consider problem domain with the *Faithfulness Condition*(Koller & Friedman, 2009). Specifically, $\mathcal{P}$ and N are faithful to each other, *iff* every conditional independence entailed by N corresponds to some Markov condition present in $\mathcal{P}$.

Due to the probabilistic nature, it is likely to find a huge number of equivalent Bayesian networks. Two different Bayesian Networks, N_1 and N_2, are independence (or Markov) equivalent, if N_1 and N_2 entail exactly the same conditional independence relations among the variables. In all these Bayesian networks, Causal Bayesian network (CBN) is a special one in which each edge is interpreted as a direct causal relationship between a parent variable node and a child variable node.

Generally speaking, it is difficult to distinguish CBN from independence equivalent Bayesian networks, unless additional assumptions are made. When the variables are correlated under linear relationships and the noises follow non-Gaussian distributions, LiNGAM and its variants (Shimizu et al., 2006; 2011) are known to return more accurate causations from the uncontrollable samples. In particular, such assumption can be formulated by an equation, such that every variable $v_i = \sum_{v_j \in P(v_i)} A_{ij} \cdot v_j + e_i$, where $P(v_i)$ contains all the parent variables of v_i, A_{ij} is the linear dependence weight w.r.t. v_i and its parent v_j, and e_i is an non-Gaussian noise over v_i. Assume that the variables in V are organized based on a topological order in the causal graph. The generation procedure of a sample could be written as $V = A \cdot V + E$. LiNGAM aims to find such a topological order and reconstructs the matrix A by exploiting *independence component analysis* (ICA) over the sample.

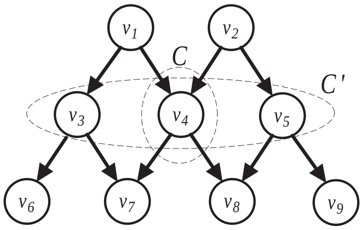


Figure 1. An example probabilistic graphical model over 9 variables and two causal cuts deduced by C and C'.

When assuming non-linear generation procedure (Hoyer et al., 2008) and discrete data domain (Peters et al., 2010), additive noise model provides another approach to utilize the asymmetric relation between causal variables and consequence variables. A regression model $v_i = f(v_j) + e_i$ is trained for each pair of variables v_i and v_j. If the noise variable e_i is independent of v_j, variable v_j is returned as the cause of variable v_i. Note that algorithms under discrete additive noise model are usually run over pairs of variables independently.

A common observation on the CBNs in real-world domains is the sparsity on the causal relationships. Specifically, a variable usually only has a small number of parental causal variables in the CBN, regardless of the underlying true generative procedure. This property, however, is not fully exploited by the existing causation algorithms.

3.2. Framework

Let $G = (V, E)$ denote a directed graph on the variable set V. A variable set $C \subset V$ forms a causal cut set over G, *iff* C deduces three non-overlapping variable subsets V_1 ,V_2 and C of V such that (1) $V_1 \cup V_2 \cup C = V$; (2) there is no edge between V_1 and V_2 in E. Intuitively, variables in C block all paths between the variables in V_1 and V_2. For each directed edge $u \rightarrow v$ in E, one of the two following cases must hold: (1) intra-causality: $u, v \in V_1$, $u, v \in V_2$ or $u, v \in C$; and (2) inter-causality: ($u \in V_1 \cup V_2$ and $v \in C$), or ($u \in C$ and $v \in V_1 \cup V_2$).

In Figure 1, for example, $C = \{v_4\}$ is a valid causal cut, which separates the variables into $V_1 = \{v_1, v_3, v_6, v_7\}$, $V_2 = \{v_2, v_5, v_8, v_9\}$. Given a directed graph G, there could be different valid cuts satisfying the above conditions. In the example graph, $C' = \{v_3, v_4, v_5\}$ is another valid causal cut with $V_1 = \{v_1, v_2\}$, $V_2 = \{v_6, v_7, v_8, v_9\}$. Note that causal cut may not lead to d-separation, e.g. C' in the exam-

Algorithm 1 SADA

```
Input: sample set $D$, variable set $V$, variable threshold $\theta$ and a causation algorithm $A$
Output: $G$: directed causal graph of the CBN
if $|V| \leq \theta$ then
    Return the result $G$ by running algorithm $A$ on $D$ and $V$.
Find a causal cut $(C, V_1, V_2)$ on $V$.
$G_1 = \mathbf{SADA}(D, V_1 \cup C, \theta, A)$.
$G_2 = \mathbf{SADA}(D, V_2 \cup C, \theta, A)$.
Return $G$ by merging $G_1$ and $G_2$.
```

Algorithm 2 Finding Causal Cut

```
Input: sample set $D$, variable set $V$
Output: a causal cut $(C, V_1, V_2)$
for $j = 1$ to $k$ do
    Randomly pick up two variables $u$ and $v$ such that $u \perp v | V - \{u, v\}$.
    Find the smallest $\hat{V} \subseteq V - u, v$ to make $u \perp v | \hat{V}$.
    Initialize $V_1 = \{u\}$, $V_2 = \{v\}$ and $C = \hat{V}$.
    Remove variables in $V_1$, $V_2$ and $C$ from $V$.
    for each variable $w \in V$ do
        if $\forall u \in V_1$, $\exists C' \subseteq C$ that $w \perp u | C'$ then
            Add $w$ into $V_2$.
        else if $\forall v \in V_2$, $\exists C' \subseteq C$ that $w \perp v | C'$ then
            Add $w$ into $V_1$.
        else
            Add $w$ into $C$.
    for each variable $s \in C$ do
        if $\forall u \in V_1$, $\exists C' \subseteq C - \{s\}$ that $s \perp u | C'$ then
            Move $s$ from $C$ to $V_2$.
        else if $\forall v \in V_2$, $\exists C' \subseteq C - \{s\}$ that $s \perp v | C'$ then
            Move $s$ from $C$ to $V_1$.
    Let $\Phi_j = (C, V_1, V_2)$
Return $\Phi_j$ with the largest $\min\{|V_1|, |V_2|\}$.
```

ple does not d-separate the other variables.

Given a causal cut (C, V_1, V_2) on variable set V, we are able to transfer the causation inference problem on V into two smaller causation inference problems over variables $V_1 \cup C$ and $V_2 \cup C$ respectively. This partitioning operation could be recursively called, until the number of variables involved in the subproblem is below a specified threshold T. The complete pseudocodes are available in Algorithm 1. The inputs of SADA include the sample set D, the variables V, a threshold θ and an underlying causation algorithm A. Here, θ is used to terminate the recursive partitioning when the variable set is sufficiently small, and A is an arbitrary causation algorithm invoked to find the actual causal graph on the subset of variables. In the rest of the section, we will discuss how to effectively and efficiently find causal cuts on a variable set V. We will also present the details of the merging operator, which tackles the problem of inconsistency and redundancy on the partial results from the subproblems.

3.3. Finding Causal Cuts

The searching of the causal cuts is crucial to the partitioning operation in SADA. To identify potential causal cuts, our algorithm resorts to conditional independence relation between variables in the Bayesian network. The following lemma formalizes the connection.

Lemma 1 (C, V_1, V_2) *is a valid causal cut over* V*, iff (1)* $V_1 \cup V_2 \cup C = V$*; and (2)* $\forall u \in V_1$ *and* $\forall v \in V_2$*, there exists a variable set* $C_{uv} \in C$ *such that* $u \perp v | C_{uv}$.

Proof: "⇒" Based on the definition of causal cut, condition (1) always holds. Since there is no directed edge between V_1 and V_2, for any $u \in V_1$and $v \in V_2$, u and v are d-separated by C, implying the validity of condition (2).

"⇐" For all pairs of variables (u, v) that $u \in V_1$ and $v \in V_2$ are d-separated by C, there is no directed edge between V_1 and V_2. Therefore, V_1, V_2 and C must satisfy the definition of causal cut. □

The details of the algorithms are listed in Algorithm 2. The algorithm runs with k different initial variable pairs. For each pair of $\{u, v\}$ conditional independent of each other in term of other variables, the algorithm greedily adds other variables into C, V_1 and V_2. After completing all assignment, the algorithm also tries to move the variables from C to V_1 or V_2 to maximize the partitioning effect. Finally, the causal cut with largest $\min\{|V_1|, |V_2|\}$ are returned as final result. We leave the discussion on the parameters k and θ to next section. Please note that the sample size needed in the cut algorithm highly depends on the local connectivity of the causal structure but not on the number of variables. This is an important advantage of the algorithm to applications in large scale sparse causal inference problems.

3.4. Merging Partial Results

As is shown in Algorithm 1, two partial results G_1 and G_2 are combined as a single casual graph as on variables in V. Since G_1 and G_2 are calculated independently, the merging operation is carefully designed to handle conflicts and redundancies.

The general form of a conflict is a cycle of directed edges among a group of variables. Given two nodes v_1 and v_2, there are two paths co-existing, such as $v_1 \to \cdots \to v_2$ and $v_1 \leftarrow v_2$. To resolve such conflicts, we simply remove the least significant edge, whenever a cycle is found.

Redundancy incurs under the following observation:

given two variables v_1 and v_2, if both $v_1 \rightarrow \cdots \rightarrow v_2$ and $v_1 \rightarrow v_2$ are discovered, $v_1 \rightarrow v_2$ may be redundant. Since the dependency relation $v_1 \rightarrow v_2$ could be blocked by certain variables in the variable set $Path(v_1 \rightarrow v_2)$, where $Path(v_1 \rightarrow v_2)$ includes all variables involved in $v_1 \rightarrow \cdots \rightarrow v_2$. Such redundancy raises when the following two conditions are satisfied: 1) the source and destination variables are both in the causal cut, i.e. $v_1, v_2 \in C$, 2) there is another variable $v_3 \in V_1$, such that $v_1 \rightarrow v_3 \rightarrow v_2$. If the above two conditions are met, one path $v_1 \rightarrow v_2$ will be returned from the subproblem with $V_1 \cup C$, while another path $v_1 \rightarrow v_2$ turns up from the other subproblem with $V_2 \cup C$. To tackle this problem, our merging algorithm runs the following conditional independence test to verify if $\exists V' \subset Path(v_1 \rightarrow v_2)$ that $v_1 \perp v_2 | Path(v_1 \rightarrow v_2)$.

To summarize, the merging operation works as follows. Firstly, all directed edges from both solutions are simply added into a single edge set. Secondly, Edges are ranked according to the associated significance measure, calculated by the underlying causation algorithm A used by SADA. Thirdly, a sequential check on the edges are run based on the order of the significance. An edge is removed if it is conflicted with any of the previous edge. Finally, the redundancy edges are discovered and removed based on results of the conditional independence testings. A complete description is available in Algorithm 3.

Algorithm 3 Merge Results

```
Input: $G_1$, $G_1$: solutions to $V_1 \cup C$ and $V_2 \cup C$
Output:$G$: solution for $V_1 \cup V_2 \cup C$
$G = G_1 \cup G_2$;
Sort edges in $G$ in descending order of significance;
Mark all variable pairs as unreachable;
for each $(v_1 \rightarrow v_2) \in G$ do
    if $(v_1, v_2)$ is reachable then
        $G = G - \{v_1 \rightarrow v_2\}$;
    else
        Mark $(v_1, v_2)$ as reachable;
for each $v_1 \rightarrow v_2 \in G$ do
    if $v_1 \rightarrow \cdots \rightarrow v_2$ is in $G$ then
        Let $Path(v_1 \rightarrow v_2)$ includes all variables involved
        in $v_1 \rightarrow \cdots \rightarrow v_2$;
        if $\exists V' \subset Path(v_1 \rightarrow v_2)$ satisfies $v_1 \perp v_2 | V'$ then
            $G = G - \{v_1 \rightarrow v_2\}$;
return $G$;
```

4. Theoretical Analysis

In this section, we study the theoretical properties of SADA, especially on the effectiveness on problem scale reduction, consistency on causal results and interpretation with independent component analysis.

4.1. Effectiveness

In this part of the section, we aim to verify the effectiveness of the causal cut search algorithm. In particular, we try to prove that the scale of the subproblem is significantly reduced when applying the randomized search algorithm.

Theorem 1 *If every variable has no more than c parental variables in CBN, by setting $k = (2c+2)^2$, Algorithm 2 returns a causal cut (C, V_1, V_2) with probability at least 0.5, such that*

$$\min\{|V_1|, |V_2|\} \geq \frac{|V|}{2c+2}$$

Proof Sketch: Since the causal graph in CBN must be a DAG, there is at least one topological order on the variables, i.e. $V = \{v_1, v_2, \ldots, v_{|V|}\}$, such that v_i's parental variables are ahead of v_i in the order. When randomly picking up variable pairs in V, i.e. u and v from V, we will first show that u and v generate a causal cut with $\min\{|V_1|, |V_2|\} \geq \frac{|v|}{2c+2}$ with probability at least $1/(2c+2)^2$.

With out loss of generality, we assume $n = |V|$ and the variable u is behind v in the topological order over V. With probability α, u is one of the variables between $v_{0.5n}$ and $v_{(0.5+\alpha)n}$. Consider all the αn variables between $v_{0.5n}$ and $v_{(0.5+\alpha)n}$. We simply put all these variables in V_1, and put all parental variables of V_1, denoted by $P(V_1)$. and all variables behind $v_{(0.5+\alpha)n}$ into C. The rest of the variables are inserted into V_2. It is easy to prove that these configuration $\{C, V_1, V_2\}$ is a valid causal cut. Moreover, $|V_1| = \alpha n$ and $|V_2| \geq \frac{n}{2} - \alpha cn$. By picking $\alpha = \frac{1}{2c+2}$, $\min\{|V_1|, |V_2|\} \geq \frac{n}{2c+2}$. When v is selected in V_2, Algorithm 2 must converge to a solution better than the artificial configuration above. This happens with probability at least $\frac{1}{(2c+2)^2}$ when $\alpha = \frac{1}{2c+2}$.

By running the randomized search algorithm $k = (2c+2)^2$ times, since $(1 - e^{-\alpha})^{\alpha} \approx e^{-1}$ when α is sufficiently large, the probability of finding a causal cut with $\min\{|V_1|, |V_2|\} \geq \frac{n}{2c+2}$ is at least 1/2. □

The last theorem implies that the causal cut is effective on reducing the scale of the subproblems. Another implication is on the selection of the parameter θ. To guarantee there is a reduction on problem size, the parameter θ should be no smaller than $2c+2$, since such theta ensuring that $\frac{\theta}{2c+2} \geq 1$.

4.2. Correctness and Completeness

The effectiveness of SADA is guaranteed based on the conclusion of the following theorem.

Theorem 2 *Assume D is a set of data samples generated from the causal structure G defined on the variables in V. If the causation algorithm A and conditional independence test used in SADA are both reliable, SADA always finds the true causal structure G.*

Proof: Assume G' is the causal structure discovered by SADA. We only need to prove the correctness and completeness of G'. The correctness and completeness are equivalent to $\forall(v_1 \to v_2) \in G'$, $(v_1 \to v_2) \in G$, and $\forall(v_1 \to v_2) \in G$, $(v_1 \to v_2) \in G'$, respectively. The details of the proof are given as follows:

Completeness: Assume $(v_1 \to v_2) \in G$, firstly, according to the causal cut step, both v_1 and v_2 must be in one subproblem, $V_1 \cup C$ or $V_2 \cup C$, but not acrose the two subproblems. Otherwise, v_1 and v_2 is conditional independent of each other given some subset of C, conflicts with the condition $v_1 \to v_2 \in G$ and the assumption that the conditional independence test is reliable. Secondly, according to the following two conditions: ”v_1 and v_2 are in the same subproblem” and ”basic causal solver is reliable”, $v_1 \to v_2 \in G'$ will be discovered in one of the subproblems. Finally, the edge $v_1 \to v_2$ won't be removed in the merge step, because if the edge is removed by either conflict or redundancy reason, it will conflict with the condition $v_1 \to v_2 \in G$ and the assumption that the condition independence test is reliable. Thus, $v_1 \to v_2$ must be contained in the result of SADA, in anther word, $v_1 \to v_2 \in G'$.

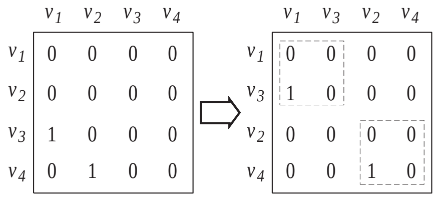
Correctness: Assume $(v_1 \to v_2) \in G'$, firstly we will show $v_1 \to v_2$ is the correct result of the subproblems. According to the framework of SADA, v_1 and v_2 must be discovered in one of the subproblem $V_1 \cup C$ and $V_2 \cup C$. Without loss of generality, assume $v_1 \to v_2$ is discovered in the subproblem $V_1 \cup C$ by the basic causal solver. According to the condition that the basic causal solver is reliable, $v_1 \to v_2$ must be the correct result for the subproblem $V_1 \cup C$. Secondly, we will show $(v_1 \to v_2) \in G$. If $v_1 \to v_2$ is the correct result of $V_1 \cup C$ but not contained in G, then there must appears some variable set $V' \subset V$ satisfies $v_1 \perp v_2 | V'$. Thus, there must be a path $v_1 \to \cdots \to v_2$ which contains V' as intermediate nodes. If such path exists, according to the Merge step, $v_1 \to v_2$ will be removed from the result set G', and conflict with the condition that $v_1 \to v_2 \in G'$. Thus, $v_1 \to v_2 \in G$. □

The previous theorem is based on an optimal setting on the reliability of the causation algorithms and conditional independence testings. In practice, such reliability may not be achieved, due to the noises on the samples and limited accuracy of the statistical numbers. In the experiments, we show that our approach remains effective, even when the condition of reliability is not fully satisfied.

4.3. Connection between SADA and ICA

In this part of the section, we discuss the connection between SADA and ICA, under the assumption of linear correlation between variables and non-Gaussian noises. By running SADA, the variables are partitioned into (possibly) overlapping subsets, such that each two subsets are independent of each other, in the sense that there is no causal edge across them. On the other hand, running ICA on the samples could be interpreted based on the causal graph represented in matrix form.

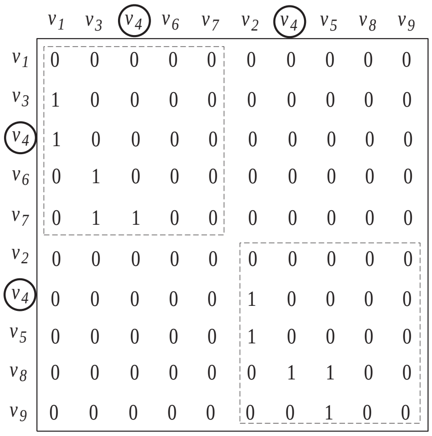
In Figure 2, we present a simple example with four variables, $\{v_1, v_2, v_3, v_4\}$, with a generative process as $v_3 = v_1 + e_3$ and $v_4 = v_2 + e_4$. An entry in the matrix indicates if there is an causal edge between two variables. Due to the sample generation rule, there are only two 1s in the matrix, i.e. (v_1, v_3) and (v_2, v_4). The ICA approach tries to find a permutation over the complete variable set, such that triangle sub-matrices with zeros on all top right entries are identified, as is shown in Figure 2. Similarly, SADA returns two subproblems based on the causal cut $C = \emptyset$, $V_1 = \{v_1, v_3\}$ and $V_2 = \{v_2, v_4\}$, by spending a much smaller computation cost.



$$\begin{array}{c|cccc} & v_1 & v_2 & v_3 & v_4 \\ \hline v_1 & 0 & 0 & 0 & 0 \\ v_2 & 0 & 0 & 0 & 0 \\ v_3 & 1 & 0 & 0 & 0 \\ v_4 & 0 & 1 & 0 & 0 \end{array} \Rightarrow \begin{array}{c|cccc} & v_1 & v_3 & v_2 & v_4 \\ \hline v_1 & 0 & 0 & 0 & 0 \\ v_3 & 1 & 0 & 0 & 0 \\ v_2 & 0 & 0 & 0 & 0 \\ v_4 & 0 & 0 & 1 & 0 \end{array}$$

Figure 2. SADA and ICA on a simple example

In a more complicated case with the causal structure in Figure 1, the causal relationships are modeled in Figure 3. Since it's impossible to find a permutation as in Figure 2 to perfectly divide the variables into triangle sub-matrices, a duplicate variable on v_4 must be introduced to satisfy the requirements of ICA. Our SADA framework easily finds the causal cut with $C = \{v_4\}$, $V_1 = \{v_1, v_3, v_6, v_7\}$ and $V_2 = \{v_2, v_5, v_8, v_9\}$, which generates subproblems consistent with ICA's results. However, the computational cost of SADA is significantly smaller than that of running ICA on the complete variable set. This is the main advantage of SADA, since much less samples are needed to find a robust causal cut and each subproblem is much easier to solve.



$$\begin{array}{c|cccccccccc}
 & v_1 & v_3 & v_4 & v_6 & v_7 & v_2 & v_4 & v_5 & v_8 & v_9 \\
\hline
v_1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
v_3 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
v_4 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
v_6 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
v_7 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
v_2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
v_4 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
v_5 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
v_8 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\
v_9 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0
\end{array}$$

Figure 3. SADA and ICA on the example in Figure 1.

5. Experiments

We evaluate our proposal on datasets generated by different real-world Bayesian network structures[1], under linear non-Gaussian model and discrete additive noise model. It generally covers a variety of applications, including, medicine (*Alarm* dataset), weather forecasting (*Hailfinder* dataset), printer troubleshooting(*Win95pts* dataset), pedigree of breeding pigs (*Pigs* dataset) and linkage among genes (*Link* dataset). The structural statistics of these Bayesian networks are summarized in Table 1. In all the Bayesian networks, the maximal degrees, i.e. the maximal number of parental variables in the networks, are no larger than 6, regardless of the total number of variables. This verifies the correctness of our sparsity assumption. On all datasets, SADA stops the partitioning when the subproblem reaches the size $\theta = 10$. The recursive partitioning is also terminated when Algorithm 2 fails to find any valid causal cut.

Table 1. Statistics on the datasets

Dataset	Variable #	Avg degree	Max degree
Alarm	37	1.2432	4
Hailfinder	56	1.1786	4
Win95pts	76	0.9211	6
Pigs	441	1.3424	2
Link	724	1.5539	3

In all the experiments, we report results on *Causal Cut Error*. We use N to denote the number of causal variable pairs in the specific Bayesian network, and use N_e to denote the number of causal variable pairs wrongly divided into subproblems after running division operations in SADA. The causal cut error is the ratio N_e/N. We also report the *recall*, *precision* and *F1 score* on the result causal relationships returned by SADA and baseline approaches. Specifically, F1 score is calculated as $\frac{2P \times R}{P+R}$, which R and P are recall and precision respectively. The experiments are compiled and run with Matlab 2009a on a windows PC equipped with a dual-core 2.93GHz CPU and 2GB RAM.

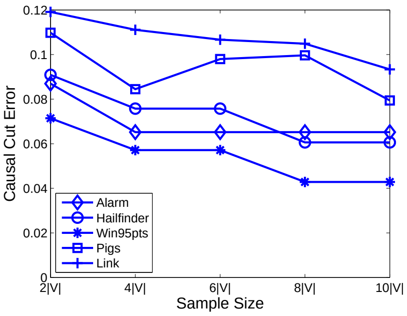


Figure 4. Causal cut errors on linear non-Gaussian models

5.1. On Linear Non-Gaussian Model

Under the assumption of linear non-Gaussian model, the samples are generated based on linear functions as $v_i = \sum_{v_j \in P(v_i)} w_{ij} v_j + e_i$. When randomly generating these linear functions, we restrict that $\sum_{P(v_i)} w_{ij} = 1$ and the the variance $Var(e_i) = 1$ for every variable v_i. We employ the conditional independence test following the method proposed in (Baba et al., 2004), with threshold at 95%. LiNGAM (Shimizu et al., 2006) is appointed as the basic causation algorithm A after SADA reaches the minimal scale threshold θ at subproblems. LiNGAM without applying any division is also used as the baseline approach, denoted by BL, when reporting recall, precision and F1 score.

The causal cut errors are reported in Figure 4, on varying the number of samples generated by the Bayesian networks. Even when the samples size is $2|V|$, the highest causal cut error is within 0.12. Moreover, the causal cut error consistently decreases with the growth of sample size. These results reveal the fundamental advantage of SADA, such that the sufficient number of samples only depends on the sparsity of the causal structure but not the number of variables. Note that the baseline approach LiNGAM does not work when the number of samples are as small as $2|V|$.

In the following experiments, we compare SADA against the baseline approach by fixing the sample size at $2|V|$. As shown in Table 2, SADA achieves significantly better F1 score on all of the five datasets. SADA is particularly doing well on precision, i.e. returning more accurate causality relationships. SADA's division strategy is the main reason behind the improvement of precision on SADA. Specifically, the division on variables allows SADA to remove a large number of candidate variable pairs if they are assigned to V_1

[1] www.cs.huji.ac.il/site/labs/compbio/Repository/

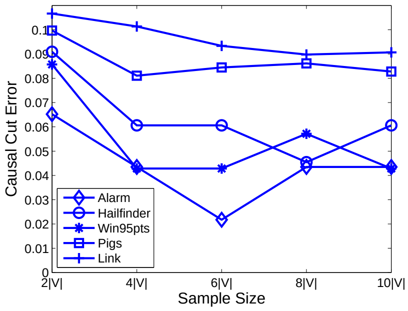


Figure 5. Causal cut errors on discrete models

and V_2. The basic causality sovler, LiNGAM in this case, is run on subproblem of much smaller scale, thus generating more reliable results. The Recall of SADA is comparable to the baseline approach on four of the datasets, and slightly worse on the other one. This shows that the unavoidable causal cut error does not affect the recall under linear non-Guassian models.

Table 2. Results on Linear Non-Gaussian Model

Dataset	Recall		Precision		F1 Score	
	SADA	BL	SADA	BL	SADA	BL
Alarm	**0.41**	0.24	**0.36**	0.30	**0.38**	0.27
Hailfinder	**0.52**	0.24	**0.46**	0.13	**0.49**	0.17
Win95pts	**0.57**	0.41	**0.42**	0.23	**0.48**	0.30
Pigs	0.56	**0.57**	**0.23**	0.12	**0.33**	0.19
Link	**0.62**	0.53	**0.25**	0.07	**0.36**	0.13

5.2. On Discrete Additive Noise Model

The generation process of the discrete data follows the method used in (Peters et al., 2011) under Additive Noise Model(ANM) for causal inference on discrete data. Each variable is restricted to 3 different value and values are randomly generated based on conditional probability tables. The implementation of SADA for discrete domain is slightly different from that for continuous domain. G^2 test (Spirtes et al., 2001) is employed as the conditional independence test, with the threshold at 95%. The causation algorithm A called by SADA is a brute force method to find all causalities on problems of small scaled. Again, the brute-force method without variable division is also employed as a baseline approach, denoted by (BL) in these results. In particular, the algorithm checks every possible pair of variables following the method proposed in (Peters et al., 2011).

The causal cut error of SADA on the discrete data is presented in Figure 5, which shows similar property of the result on linear non-Gaussian models. This further verifies the generality of SADA on different data domains.

In this group of experiments, we fix the sample size at 2000, and report recall, precision and F1 score in Table 3. Note that the baseline approach is only applicable to domain with small number of variables. This leads to difficulties for baseline to finish the computation on *Pigs* and *Link* in one week. This proves the improvement of SADA on scalability in terms of the variables. Generally speaking, the results in the table also verifies the effectiveness of SADA, especially on dramatic enhancement on precision and F1 score.

Table 3. Results on Discrete Model

Dataset	Recall		Precision		F1 Score	
	SADA	BL	SADA	BL	SADA	BL
Alarm	**0.67**	0.65	**0.72**	0.60	**0.70**	0.63
Hailfinder	0.71	**0.76**	**0.57**	0.45	**0.63**	0.56
Win95pts	0.68	**0.71**	**0.41**	0.38	**0.51**	0.49
Pigs	**0.68**	N.A.	**0.50**	N.A.	**0.58**	N.A.
Link	**0.69**	N.A.	**0.46**	N.A.	**0.56**	N.A.

As a conclusion, SADA shows excellent performance on 5 different domains with real-world Bayesian networks. SADA returns accurate causal relations when combined with two well known causal inference algorithms. The causal cut used to partition the problem does incur certain error on incorrect partitioning. Despite of the errors, SADA still outperforms the baseline algorithms without partitioning on almost all settings.

6. Conclusion

In this paper, we present a general and scalable framework, called SADA, to support causal structure inference, using a *split-and-merge* strategy. In SADA, causal inference problem on a large variable set is partitioned into subproblems with overlapping subsets of variables, utilizing the concept of causal cut. Our proposal facilitates existing causation algorithms to handle problem domains with more variables and less samples, which are considered impossible in the past. Strong theoretical analysis proves the effectiveness, correctness and completeness guarantee of SADA under a general setting. Experimental results further verifies the usefulness of the new framework with two mainstream causation algorithms on linear non-Gaussian model and discrete additive noise model.

7. Acknowledgements

Cai and Hao are supported by NSF of China (61070033, 61100148,61202269), NSF of Guangdong Province (S2011040004804), Opening Project of the State Key Laboratory for Novel Software Technology (KFKT2011B19), Foundation for Distinguished Young Talents in Higher Education of Guangdong, China (LYM11060).

References

Aliferis, Constantin F., Statnikov, Alexander, Tsamardinos, Ioannis, Mani, Subramani, and Koutsoukos, Xenofon D. Local causal and markov blanket induction for causal discovery and feature selection for classification. *J. Mach. Learn. Res.*, 11:171–234, 2010.

Baba, Kunihiro, Shibata, Ritei, and Sibuya, Masaaki. Partial correlation and conditional correlation as measures of conditional independence. *Australian & New Zealand Journal of Statistics*, 46(4):657–664, 2004.

Bromberg, Facundo and Margaritis, Dimitris. Improving the reliability of causal discovery from small data sets using argumentation. *J. Mach. Learn. Res.*, 10: 301–340, 2009.

Cai, Ruichu, Zhang, Zhenjie, and Hao, Zhifeng. Bassum: A bayesian semi-supervised method for classification feature selection. *Pattern Recognition*, 44 (4):811–820, 2011.

Cai, Ruichu, Zhang, Zhenjie, and Hao, Zhifeng. Causal gene identification using combinatorial v-structure search. *Neural Networks*, 2013. doi: 10.1016/j.neunet.2013.01.025.

Friedman, Nir, Linial, Michal, Nachman, Iftach, and Pe'er, Dana. Using bayesian networks to analyze expression data. In *RECOMB*, pp. 127–135, 2000.

He, Y. and Geng, Z. Active learning of causal networks with intervention experiments and optimal designs. *J. Mach. Learn. Res.*, 9:2523C2547, 2008.

Hoyer, Patrik O., Janzing, Dominik, Mooij, Joris M., Peters, Jonas, and Schölkopf, Bernhard. Nonlinear causal discovery with additive noise models. In *NIPS*, pp. 689–696, 2008.

Janzing, Dominik, Mooij, Joris M., Zhang, Kun, Lemeire, Jan, Zscheischler, Jakob, Daniusis, Povilas, Steudel, Bastian, and Schölkopf, Bernhard. Information-geometric approach to inferring causal directions. *Artif. Intell.*, 182-183:1–31, 2012.

Kalisch, M. and Bühlmann, P. Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *The J. Mach. Learn. Res.*, 8:613–636, 2007.

Kim, Sunyong, Imoto, Seiya, and Miyano, Satoru. Dynamic bayesian network and nonparametric regression for nonlinear modeling of gene networks from time series gene expression data. *Biosystems*, 75(1-3):57–65, 2004.

Koller, Daphne and Friedman, Nir. *Probabilistic Graphical Model: Principles and Techniques*. The MIT Press, 2 edition, 2009.

Mooij, Joris M., Janzing, Dominik, Peters, Jonas, and Schölkopf, Bernhard. Regression by dependence minimization and its application to causal inference in additive noise models. In *ICML*, pp. 94, 2009.

Pearl, Judea. *Causality: models, reasoning and inference*. Cambridge Univ. Press, 2 edition, 2009.

Pearl, Judea and Verma, Thomas. A theory of inferred causation. In *Proceedings of the 2nd International Conference on Principles of Knowledge Representation and Reasoning*, pp. 441–452, 1991.

Peters, Jonas, Janzing, Dominik, and Schölkopf, Bernhard. Identifying cause and effect on discrete data using additive noise models. In *AIStats*, pp. 597–604, 2010.

Peters, Jonas, Janzing, Dominik, and Schölkopf, Bernhard. Causal inference on discrete data using additive noise models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(12):2436 – 2450, 2011.

Shimizu, Shohei, Hoyer, Patrik O., Hyvärinen, Aapo, and Kerminen, Antti J. A linear non-gaussian acyclic model for causal discovery. *J. Mach. Learn. Res.*, 7:2003–2030, 2006.

Shimizu, Shohei, Inazumi, Takanori, Sogawa, Yasuhiro, Hyvärinen, Aapo, Kawahara, Yoshinobu, Washio, Takashi, Hoyer, Patrik O., and Bollen, Kenneth. Directlingam: A direct method for learning a linear non-gaussian structural equation model. *J. Mach. Learn. Res.*, 12:1225–1248, 2011.

Spirtes, Peter, Glymour, Clark, and Scheines, Richard. *Causation, Prediction, and Search*. The MIT Press, 2 edition, 2001.

Zhang, Kun, Peters, Jonas, Janzing, Dominik, and Schölkopf, Bernhard. Kernel-based conditional independence test and application in causal discovery. *CoRR*, abs/1202.3775, 2012.

Zhu, Z., Ong, Y.S., and Dash, M. Markov blanket-embedded genetic algorithm for gene selection. *Pattern Recognition*, 40(11):3236–3248, 2007.

Zscheischler, Jakob, Janzing, Dominik, and Zhang, Kun. Testing whether linear equations are causal: A free probability theory approach. *CoRR*, abs/1202.3779, 2012.