Causal query in observational data with hidden variables

Debo Cheng¹, Jiuyong Li¹, Lin Liu¹, Jixue Liu¹, Kui Yu², and Thuc Duy Le¹

Abstract. This paper discusses the problem of causal query in observational data with hidden variables, with the aim of seeking the change of an outcome when “manipulating” a variable while given a set of plausible confounding variables which affect the manipulated variable and the outcome. Such an “experiment on data” to estimate the causal effect of the manipulated variable is useful for validating an experiment design using historical data or for exploring confounders when studying a new relationship. However, existing data-driven methods for causal effect estimation face some major challenges, including poor scalability with high dimensional data, low estimation accuracy due to heuristics used by the global causal structure learning algorithms, and the assumption of causal sufficiency when hidden variables are inevitable in data. In this paper, we develop theorems for using local search to find a superset of the adjustment (or confounding) variables for causal effect estimation from observational data under a realistic pretreatment assumption. The theorems ensure that the unbiased estimate of causal effect is included in the set of causal effects estimated by the superset of adjustment variables. Based on the developed theorems, we propose a data-driven algorithm for causal query. Experiments show that the proposed algorithm is faster and produces better causal effect estimation than existing data-driven causal effect estimation methods with hidden variables. The causal effects estimated by the proposed algorithm are as accurate as those by the state-of-the-art methods using domain knowledge.

1 Introduction

Data is frequently used for various decision making which involves causal evidence seeking (or causal query) in observational data [25, 20, 35, 40]. For example, a biologist is planning for a biological experiment, and she has a collection of data from previous experiments by other researchers (those experiments do not need to have the same objective as hers) [6, 19]. She can plan the experiment on the data by “manipulating” the variable which is to be modified in the planned experiment, i.e. the treatment variable to see if the outcome is changed as expected. This type of query is useful in many real-world applications. When a public policy is in the design phase, existing data can be used to assess the policy to find out whether the policy would produce the desired outcome and what major factors would affect the outcome. In a marketing campaign, similar queries can be made to a dataset to help plan a successful campaign.

In this paper, we consider the problem that a user queries a dataset for a causal effect, i.e. how much the interested outcome Y will change due to a change of the treatment W. The answer depends on what other variables affect W and Y, called confounders or adjustment variables. Users do not know the adjustment variables in most cases. For example, when a biologist studies how a gene causes cancer, she may not know (or only partially know) the other genes which are causal factors of the cancer and regulate this gene under study. So, the adjustment variables are a part of the answer to seek too and the query answer will be a set of pairs each containing a set of possible adjustment variables (called an adjustment set in the paper) and the corresponding causal effect. To confirm which pair corresponds to the real-world mechanism, a user can design an experiment by physically controlling the variables in an adjustment set while manipulating the treatment variable, or pick up one or multiple returned adjustment sets consistent with her knowledge of the system as plausible answers.

From data, especially data with hidden variables [5, 27, 39], it may not be possible to find an exact adjustment set. The number of returned possible adjustment sets can be very large, which will make the query answer unusable. For example, a biologist has only a limited budget to try a few experiments, and she will not be able to try many possible adjustment sets. An advertisement company can try a limited number of A/B tests to make a promotion decision. Another constraint is the size of an adjustment set. A long adjustment set containing many variables will complicate the design of an experiment and make it hard to explain experimental results.

The causal query problem studied in this paper is related to adjustment variable search and causal effect estimation in data. However, existing works on causal effect estimation often have an adjustment set given [16, 21, 31]. Hence they are not useful to solve our problem.

Graphical causal modelling is a principled approach to adjustment set search in data and causal effect estimation. Pearl has proposed the back-door criterion for identifying an adjustment set based on a given causal DAG (direct acyclic graph), and the do calculus for deriving identifiable causal effects from the given causal DAG and data [25]. The back-door criterion has become the main principle for identifying adjustment sets.

When there are no hidden variables or more precisely the causal sufficiency assumption is satisfied, i.e. all common causes are observed [32], we can learn from data a Markov equivalence class of causal DAGs for causal effect estimation. IDA is an algorithm for estimating causal effects directly from data [23]. As many equivalent causal DAGs can be learned from a dataset, IDA returns a multiset of causal effects with different adjustment sets blocking the back-door paths into the treatment variable W. Recently, Johansson et al. utilised deep learning to learn balanced representations for counterfactual inference [16]. Furthermore, Shalit et al. gave a meaningful and intuitive error bound to guide deep neural networks for estimating individual causal effects [31]. However, the algorithms based on

¹ School of Information Technology and Mathematical Sciences, University of South Australia, Mawson Lakes, SA 5095, Australia. E-mail: chedy055@mymail.unisa.edu.au and Jiuyong.Li@unisa.edu.au.
² School of Computer Science and Information Engineering, Hefei University of Technology, Hefei, China.
deep learning do not provide an explicit adjustment variable set.

When there are hidden variables, LV-IDA \cite{24} extends IDA based to a MAG (Maximal Ancestral Graph) which represents a causal structure with hidden variables. LV-IDA also returns a multiset of causal effects. Furthermore, LV-IDA has low efficiency for high-dimensional or large data. Hyttinen et al. presented an ASP constraint solver for causal inference directly in data without assuming causal sufficiency \cite{13}, but it is impractical for high-dimensional data due to the fact that the query-based techniques need to perform a complete search for solutions. Entner et al. proposed a data-driven approach to estimating causal effect from data with hidden variables \cite{10}, but with impractically high time complexity. Louizos et al. developed a new neural network latent variables model to estimate population causal effects \cite{21}, but the model relies on a correct proxy variable estimation and does not provide an explicit adjustment set.

In this paper, we aim to develop a practical solution for efficient and accurate causal query in data without assuming causal sufficiency. The contributions of the work are:

- We have developed the theorems for finding a superset of adjustment variables for unbiased causal effect estimation based on the local causal structure learning in a MAG under a realistic pre-treatment assumption. The theorems support efficient data-driven causal effect estimations without assuming causal sufficiency in data. To our best knowledge, they are the first theorems to support such a search.

- We have proposed an efficient and effective Data-driven Causal Effect estimation (DICE) algorithm. DICE is faster than LV-IDA, the only practical data-driven causal effect estimation method with hidden variables in graphical causal modelling. Extensive experiments show that DICE works in datasets where LV-IDA fails, and produces better causal effect estimation than LV-IDA. The causal effects estimated by DICE are as accurate as those by the state-of-the-art methods using domain knowledge.

2 Preliminaries

2.1 Notation and basic definitions

We use upper case letters to represent variables and bold-faced upper case letters to denote sets of variables. Let $G = (V, E)$ be a graph, where $V = \{V_1, ..., V_n\}$ is the set of nodes and $E$ a set of edges.

Two nodes are adjacent if there is an edge between them. The set of all the nodes adjacent to node $V$ is denoted as $Adj(V)$. A path $\pi$ from $V_i$ to $V_j$ is a sequence of distinct nodes $< V_{i_1}, ..., V_{i_k} >$ such that every pair of successive nodes are adjacent in $G$. A sub-path of $\pi$ from $V_i$ to $V_j$ is denoted by $\pi(V_i, V_j)$. In $G$, if there exists $V_i \rightarrow V_j$, $V_i$ is a parent of $V_j$ and we use $Pa(V_j)$ to denote the set of all parents of $V_j$. If there exists $V_i \leftarrow V_j$, $V_i$ is a spouse of $V_j$ and we use $Sp(V_j)$ to denote the set of all spouses of $V_j$. A path from $V_i$ to $V_j$ is a directed path if for any pair of adjacent nodes $V_i$ and $V_j$ in the path where $i \neq j$, $V_i$ is a parent of $V_j$, $V_j$ is an ancestor of $V_i$, and there is a directed path from $V_i$ to $V_j$. The ancestor set of $V_i$ is denoted as $An(V_i)$.

A DAG is a directed graph without directed cycles (See an example of directed cycle in Figure 1 (a)). When a DAG satisfies the following Markov condition and faithfulness assumption, we can read dependency/independency in a distribution in DAG.

**Definition 2.1** (Markov condition \cite{25}). Given a DAG $G = (V, E)$ and $P(V)$, the joint probability distribution of $V$, $G$ satisfies the Markov condition if for $\forall V_i \in V$, $Vi$ probabilistically independent of all non-descendants of $V_i$, given the parents of $V_i$.

Based on the Markov condition, according to $G P(V)$ can be factorized into: $P(V) = \prod_{i} P(V_i | Pa(V_i))$.

**Definition 2.2** (Faithfulness \cite{32}). A DAG $G = (V, E)$ is faithful to $P(V)$ if every independence presenting in $P(V)$ is entailed by $G$ and fulfills the Markov condition. A distribution $P(V)$ is faithful to a DAG $G$ iff there exists a DAG $G$ which is faithful to $P(V)$.

A causal DAG is a DAG in which a node's parents are interpreted as its direct causes. To learn a causal DAG from observational data, we also need to assume causal sufficiency.

**Definition 2.3** (Causal sufficiency \cite{32}). A dataset satisfies causal sufficiency if for every pair of variables $(V_i, V_j)$ in $V$, all their common causes are also in $V$.

When causal sufficiency is violated, ancestral graphs are used to represent data generating processes. An ancestral graph $\mathcal{M}$ is a mixed graph that does not contain directed cycles or almost directed cycles. A graph is called a mixed graph if it includes directed and bi-directed edges. When there exists a bi-directed edge $V_i \leftrightarrow V_j$ in $\mathcal{M}$, there is not directed path $V_i \rightarrow V_j$ or $V_i \leftarrow V_j$. An almost directed cycle occurs when $V_i \leftrightarrow V_j$ is in $\mathcal{M}$ and $V_j \in An(V_i)$. Figure 1 (b) is such an example. Let "∗" denote any allowed edge marks. For a path $\pi$, a non-ending node $V_i$ is a collider on $\pi$ if $\pi$ contains $V_{i-1} \rightarrow V_i \leftrightarrow V_j \rightarrow V_{i+1}$. A path $\pi$ is a collider path if each node excluding the ending nodes on $\pi$ is a collider.

**Definition 2.4** (m-separation \cite{27}). In an ancestral graph $\mathcal{G}$, a path $\pi$ between $V_i$ and $V_j$ is said to be m-separated by a set of nodes $Z$ (possibly $\emptyset$) if (1). $\pi$ does not contain any collider which is in $Z$, or (2). for any collider $V_k$ on the path $\pi$, $V_k \notin Z$ and no descendant of $V_k$ is in $Z$. Two nodes $V_i$ and $V_j$ are said to be m-connected by $Z$ in $\mathcal{G}$ if $V_i$ and $V_j$ are not m-separated by $Z$.

If $V_i$ and $V_j$ are m-separated by $Z$, $V_i \perp \!\!\!\perp V_j | Z$, the information flow from $V_i$ and $V_j$ is "blocked" by $Z$. In this paper, we call $Z$ a block set of the ordered pair ($V_i, V_j$).

**Definition 2.5** (Maximal ancestral graph (MAG)). An ancestral graph $\mathcal{M} = (V, E)$ is called a maximal ancestral graph if every pair of non-adjacent nodes $V_i$, $V_j$ in $\mathcal{M}$ can be m-separated by a set $Z \subseteq V \setminus \{V_i, V_j\}$.

We present an example of causal MAG in Figure 2 (b). Figure 2 (a) is its corresponding causal DAG with the hidden variables $U_1$ and $U_2$ shown. In a MAG $\mathcal{M}$, if a path between $V_i$ and $V_j$ is a directed path, it is called a causal path from $V_i$ to $V_j$. A non-directed path between $V_i$ and $V_j$ is a non-causal path. In Figure 2 (b), path $W \rightarrow Y$ is a causal path, and other paths between $W$ and $Y$ are non-causal paths.

**Definition 2.6** (Visibility \cite{41}). Given a MAG $\mathcal{M}$, a directed edge $V_i \rightarrow V_j$ is visible if there is a node $V_k$ not adjacent to $V_j$, such that either there is an edge between $V_i$ and $V_k$, that is into $V_i$, or there is a collider path between $V_i$ and $V_k$, that is into $V_i$ and every node in this path is a parent of $V_k$. Otherwise, $V_i \rightarrow V_j$ is said to be invisible.
Figure 2. (a) A DAG with two hidden variables $U_1$ and $U_2$. (b) The corresponding MAG of the DAG.

Figure 3. Two possible configurations of the visible edge for $V_i \rightarrow V_j$.

Let $W$ be a binary treatment variable, $Y$ the outcome variable. The causal effect of $W$ on $Y$, $CE(W, Y)$ is the change of $Y$ due to a change of $W$. When we use the do operator \[\text{do}(w)\] to represent a manipulation of a variable (e.g. set $W$ to 1) and denote $W = 1$ and $W = 0$ as $w, w'$ respectively, $CE(W, Y)$ is defined as:

$$CE(W, Y) = E(Y \mid \text{do}(w)) - E(Y \mid \text{do}(w')) \quad (1)$$

where $E$ is the Expectation. $CE(W, Y)$ can be found in an experimental setting where $W$ is physically set as 0 and 1, respectively. Let us assume that an underlying mechanism dictates the causal relationships among variables, and $CE(W, Y)$ is determined by the mechanism.

When estimating the causal effect of $W$ on $Y$, the effect of other variables needs to be eliminated or adjusted to obtain an unbiased estimation of the causal effect. In graphical causal modelling, the back-door criterion is the main principle for identifying a set of adjustment variables, denoted as $Z$ in this paper.

**Definition 2.7** (Back-door criterion in DAG). A set of variables $Z$ satisfies the back-door criterion relative to $(W, Y)$ in a DAG $G$ if (1) $Z$ does not contain descendants of $W$; (2) $Z$ blocks every back-door path between $W$ and $Y$, (i.e. the paths with an arrow pointing to $W$).

Given an adjustment set $Z$, the causal effect $CE(W, Y)$ can be estimated unbiasedly as follows:

$$CE(W, Y) = \sum_{Z} [E(Y \mid w, Z = z) - E(Y \mid w', Z = z)]p(Z = z) \quad (2)$$

When the assumption of causal sufficiency does not hold, the generalised back-door criterion in a MAG can be used to search for an appropriate adjustment set $Z$ to estimate $CE(W, Y)$ \[\text{[22]}.\]

**Definition 2.8** (Back-door path in MAG). For the ordered pair $(W, Y)$ in a MAG, a path from $W$ to $Y$ is a back-door path if it does not have a visible edge out of $W$.

**Definition 2.9** (Generalised back-door criterion for MAG). Given a MAG $M = (V, E)$, a set $Z \subseteq V \setminus \{W, Y\}$ satisfies the generalised back-door criterion w.r.t. $(W, Y)$ in $M$ if (1) $Z$ does not contain descendants of $W$; (2) $Z$ blocks every back-door path between $W$ and $Y$.

If a set $Z$ in $M$ satisfies the generalised back-door criterion, relative to $(W, Y)$, then the unbiased estimation of $CE(W, Y)$ can be achieved by Eq. \[\text{[8]}\].

In some MAGs, the adjustment sets could not be found due to the causal ambiguities, and this can be judged based on the amenability of a MAG \[\text{[33]}.\]

**Definition 2.10** (Amenable MAG w.r.t. $(W, Y)$). Let $W$ and $Y$ be two nodes in a MAG $M = (V, E)$. The MAG is adjustment amenable w.r.t. $(W, Y)$ if $W \rightarrow Y$ is visible.

If the MAG $M$ is not adjustment amenable w.r.t. $(W, Y)$, then no adjustment set $Z$ in $V \setminus \{W, Y\}$ can be found \[\text{[33]}\].

3 Local search for an adjustment set

3.1 The Theorems

Let $D$ be a dataset containing a binary treatment variable $W$, an outcome variable $Y$ which is binary or numerical, and $X$, a set of all other variables of any type. We assume that the dataset is generated from an underlying causal MAG $M$ which is adjustment amenable to the ordered pair $(W, Y)$. We also assume that all the variables in $X$ are measured before applying the treatment and observing $Y$, indicating that variables in $X$ are all non-descendants of $W$ or $Y$ in $M$. The pretreatment variable assumption is realistic as it reflects what normally happens in practice, that is, all the other variables are often measured before the treatment is applied and the outcome is observed. This assumption is commonly used in \[\text{[9, 12, 14, 36]}\].

In the proposed Theorem below, $M_{\text{MIP}}$ denotes the manipulated $M$ from which $W \rightarrow Y$ has been removed, and $M_{\text{MIP}}$ is still a MAG by the closure property of a MAG \[\text{[27]}\].

**Theorem 1** (Adjustment set discovery via local search). Given a causal MAG $M$ which is adjustment amenable to the ordered pair $(W, Y)$, there exists at least an adjustment set $Z \subseteq X$ for $(W, Y)$ such that $Z$ is a subset of $\text{Adj}(W \cup Y)$, where $\text{Adj}(W \cup Y)$ is the shorthand of $\text{Adj}(W) \cup \text{Adj}(Y)$ in $M_{\text{MIP}}$.

**Proof.** The requirement that $M$ is adjustment amenable to $W$ and $Y$ is to ensure the identifiability of the causal effect of $W$ on $Y$. A set $Z$ blocking all back-door paths from $W$ to $Y$ is an adjustment set for the ordered pair $(W, Y)$. There are two types of paths between two variables, causal paths and non-causal paths. In our problem setting, pretreatment variables $X$ are non-descendants of $W$ and $Y$ and hence $M_{\text{MIP}}$ does not contain any causal path between $W$ and $Y$.

To prove that an adjustment set $Z$ in the MAG $M$ is a subset of $\text{Adj}(W \cup Y)$, we will show that any non-causal path from $W$ to $Y$ is blocked by a variable in $\text{Adj}(W \cup Y) = \text{Pa}(W \cup Y) \cup \text{Sp}(W \cup Y)$ where $\text{Pa}(W \cup Y)$ denotes the set of $W$’s parents and $Y$’s parents, and $\text{Sp}(W \cup Y)$ denotes the set of $W$’s spouses and $Y$’s spouses. There are four cases of a non-causal path from $W$ to $Y$ as follows.

1. No colliders are in $\text{Pa}(W \cup Y)$. Each path will be blocked by a node either in $\text{Pa}(W)$ or $\text{Pa}(Y)$, or the empty set when the path contains one or more colliders.

2. If a node on a path is in $\text{Sp}(W \cup Y)$ (denoted as $S$). 2a. If $S$ is a collider, the path including $S$ is blocked by the empty set. 2b.
If $S$ is not a collider, the path is in the form of $W \leftrightarrow S \rightarrow \ldots$ or $\ldots \leftarrow S \leftarrow Y$, then $S$ blocks the path or the empty set blocks the path if the path includes one or more colliders. A special case, $W \leftarrow \ldots \leftarrow S \leftarrow Y \leftarrow W$, forms an almost cycle, which is not allowed in a MAG, so we don’t consider it.

3. If two nodes on a path $\pi$ are both in $Sp(W \cup Y)$ (denoted as $S_1$ and $S_2$). If $S_1 \neq S_2$, then the path will be blocked as in Case 2.

4. If a node $S' \in Adj(W \cup Y)$ on the path $\pi$ is a descendant of a collider $C$ in another path $\pi'$. The path $\pi$ has to be blocked by $S'$ (no other alternatives). As in Case 1, the path $\pi'$ is not opened since a parent of $W$ or $Y$ will still block it. When $C$ is a blocking variable as in Cases 2 & 3a, one end of the path $\pi'$ to $W$ or $Y$ is opened, but the other end can be blocked. As in Case 3b, an almost cycle is formed, but since it is not allowed by a MAG, this case is impossible.

Therefore, there exists at least an adjustment set $Z \subseteq Adj(W \cup Y)$ for the ordered pair $(W, Y)$ in the causal MAG $M$.

We can refine $Adj(W \cup Y)$ further according to the following theorem.

**Theorem 2** (Refinement of the superset of adjustment set). Given a causal MAG $M$ which is adjustment amenable to the ordered pair $(W, Y)$, there exists at least an adjustment set $Z \subseteq X$ for $(W, Y)$ such that $Z$ is a subset of $\text{Adj}(W \cup Y) = \text{Adj}(W \cup Y) \setminus S$, where each $X \in S$ is either $m$-separated from $Y$ by $W$ or $m$-separated from $W$ by the empty set.

**Proof.** If $X$ is $m$-separated from $Y$ by $W$, then $X$ and $Y$ are not $m$-connecting given $W$. Hence we can remove $X$ from an adjustment set. If $X$ is $m$-separated from the empty set, then $X$ and $W$ are not $m$-connecting given the empty set. Hence we can remove $X$ from an adjustment set.

### 3.2 Causal effect estimation

Based on the above two Theorems developed, we can firstly employ a local causal structure discovery algorithm to obtain $\text{Adj}_R(W \cup Y)$ from data, then our answer to a causal query is an Adjustment Set-Causal Effect Table, or ASCET, where each row contains a candidate adjustment set (a subset of $\text{Adj}_R(W \cup Y)$) and the causal effect estimated corresponding to the candidate adjustment set.

**Example.** Suppose $\text{Adj}_R(W \cup Y) = \{X_1, X_2, X_3\}$, Table 1 shows an example ASCET where 1 denotes a variable is in a candidate adjustment set and 0 otherwise.

### 3.3 Removing insignificant adjustment variables

We reduce the size of ASCET by removing variables that do not significantly affect causal effect estimation. To test if a variable significantly affects causal effect estimation, we compare the cause effect estimated by including the variable with the cause effect estimated by excluding the variable in all cases. If the average difference is smaller than a threshold, the variable is insignificant. More precisely, we have the following definition.

**Definition 3.1.** For each $X \in \text{Adj}_R(W \cup Y)$, let $Z \subseteq \text{Adj}_R(W \cup Y)$ be a candidate adjustment set containing $X$ and $Z' = Z \setminus \{X\}$, the sensitivity of $X$ is defined as:

$$\text{Sen}(X) = \frac{1}{p} \sum_{Z} |CE_Z - CE_{Z'}|$$

where $p$ is the size of the power set of $\text{Adj}_R(W \cup Y) \setminus \{X\}$, $CE_Z$ and $CE_{Z'}$ are average causal effects of $W$ on $Y$ with the adjustment sets $Z$ and $Z'$ respectively.

The sensitivity of variable $X$ reflects its impact on the estimated causal effect of $W$ on $Y$. When $X$ has a low sensitivity, no significant error will be introduced if we exclude it from $\text{Adj}_R(W \cup Y)$. So, $X$ is removed from $\text{Adj}_R(W \cup Y)$, and the size of ASCET is reduced.

**Example.** We use the example in Table 1 to illustrate how we reduce the size of ASCET. We use $X_1$ as an example, $p = 4$ as the size of the power set of $\{X_2, X_3\}$ is 4, then $\text{Sen}(X_1) = \frac{1}{4}(|CE_{X_1} - CE_{X_2}| + |CE_{X_1} - CE_{X_3}| + |CE_{X_1} - CE_{X_1}| + |CE_{X_1} - CE_{X_1}| + |CE_{X_1} - CE_{X_1} - CE_{X_2}|) = \frac{1}{4}(0.1 + 0.4 + 0.4 + 0.4) = 0.25$.

In the same way, $\text{Sen}(X_2) = 0.025$ and $\text{Sen}(X_3) = 0.175$. Hence, $X_2$ has a low sensitivity such that we can exclude it from $\text{Adj}_R(W \cup Y)$ and the corresponding rows in ASCET can be removed. The final result is presented as Table 2.

### 3.4 The proposed DICE algorithm

Algorithm 1 presents our proposed algorithm for Data-driven Causal Effect estimation without causal sufficiency (DICE). DICE contains two parts. Part 1 (lines 1 to 13) is for finding $\text{Adj}_R(W \cup Y)$. Firstly, a local causal structure learning algorithm (such as PC-Select [4] or HITON-PC [1]) is used to find the parents of $W$ and $Y$. In this work, we use PC-Select, given its high accuracy, PC-Select is a local version of PC [32], an algorithm for learning a global causal structure. Then we check whether $\text{Adj}_R(W) = \text{Adj}_R(W) \setminus \{X\}$ is empty to determine if the causal effect of $W$ on $Y$ is identifiable or not from data. If $\text{Adj}(W)$ is empty, no adjustment set can be found, and DICE returns an empty ASCET and terminates; Otherwise, we refine $\text{Adj}_R(W \cup Y)$ to achieve $\text{Adj}_R(W \cup Y)$.

Part 2 of Algorithm 1 (lines 14 to 23) firstly estimates each possible causal effect when adjusting each subset of $\text{Adj}_R(W \cup Y)$. We use Propensity Score Matching (PSM) [33] to estimate causal effects where propensity score is calculated by glm and matching is performed by Match in the R packages stats and Matching [34], respectively. The adjustment sets and the corresponding causal effects estimated are added to the ASCET. Next, DICE calculates the sensitivity of each variable in $\text{Adj}_R(W \cup Y)$, and if a variable’s sensitivity is below the given threshold, DICE removes all the adjustment sets containing the variable and the corresponding causal effects from the ASCET.
Algorithm 1 Data-driven Causal Effect estimation without causal sufficiency (DICE)

Input: Dataset $D$ with treatment $W$, pretreatment variables $X$, outcome $Y$ and a sensitivity threshold $\tau$.

Output: The ASCET.

1. Call a local causal structure learning algorithm to find $\text{Adj}(W) \setminus \{Y\}$ and $\text{Adj}(Y) \setminus \{W\}$ from $D$, respectively.
2. $\text{ASCET} \leftarrow \emptyset$
3. if $\text{Adj}(W)$ is empty then
4. Adjustment set could not be found.
5. return $\text{ASCET}$.
6. else
7. $S \leftarrow \emptyset$
8. for each variable $X \in \text{Adj}(W \cup Y)$ do
9. if $X \perp \perp W \mid \emptyset$ or $X \perp \perp Y \mid W$ then
10. add $X$ to $S$
11. end if
12. end for
13. $\text{Adj}_R(W \cup Y) = \text{Adj}(W \cup Y) \setminus S$
14. for each subset $Z \subseteq \text{Adj}_R(W \cup Y)$ do
15. calculate $\text{CE}(W, Y)$ by adjusting $Z$.
16. add pair $(Z, \text{CE}(W, Y))$ to ASCET.
17. end for
18. For each $X \in \text{Adj}_R(W \cup Y)$ calculate $\text{Sen}(X)$.
19. if $\text{Sen}(X) < \tau$ then
20. remove from ASCET all the rows containing $X$.
21. end if
22. return $\text{ASCET}$.
23. end if

Time-complexity analysis. PC-Select and PSM contribute the most to the time complexity of DICE. In the worst case, the complexity of PC-Select is $O(p2^{p-n})$ [4], where $p$ is the number of variables and $n$ the number of samples. In our problem, the worst case is when all variables of $X$ are causes of $W$ and/or $Y$, which, however rarely occurs. In most cases, PC-Select can handle thousands of variables [4]. The time complexity of PSM is $O(p^2)$. Our experiment in Section 4 shows that DICE scales with both $p$ and $n$ well.

4 Experiments

4.1 The quality of local structure learning

Before evaluating DICE, we assess the quality of local structure learning algorithm, i.e. PC-Select used in our paper versus global structure learning algorithms. The synthetic datasets are generated according to the underlying causal DAG in Fig 2(a) with 8 pretreatment variables, 2 hidden variables, $W$ and $Y$ by following the procedure in [11]. The underlying causal MAG is presented in Fig 2(b), i.e. $U_1$ and $U_2$ are removed from these generated datasets.

We generate two groups of datasets to evaluate the impact of sample size and the number of variables on the quality of the local structure learning algorithm, respectively. The first group of datasets is 10 datasets, each containing 5K samples for 10 variables generated from Fig 2(b), plus 0, 10, 20, . . . , 90 random variables, respectively. The second group of datasets include 20 datasets with 10 fixed variables of the MAG, but with varying sample sizes of 5K, 10K, . . . , 95K and 100K. These datasets with hidden variables are used to evaluate the quality of the local structure learning algorithm.

Figure 4. The Precision, Recall and F-score of FCI & RFCI & PC-Select algorithms on two groups of synthetic datasets w.r.t. the number of variables and samples. The qualities of parent discovery by the three algorithms are very consistent, but the time efficiencies of the algorithms are quite different as shown in Figure 5.

We use two global structure learning algorithms, Fast causal inference (FCI) [32] and Really FCI (RFCI) [8] to learn from each of the generated datasets a PAG (Partial Ancestral Graph) which represents a Markov equivalence class of MAGs encoding the same dependence/independence relations in the data. Then we extract $\text{Adj}(W)$ and $\text{Adj}(Y)$ from the learned PAG. For PC-Select, we apply it twice to a dataset, one having $W$ as the target and the other having $Y$ as the target to learn $\text{Adj}(W)$ and $\text{Adj}(Y)$. The implementations of FCI, RFCI and PC-Select are from the R package pcalg [17], with the default parameter settings and a significance level ($\alpha$) of 0.05.

Results. We perform the experiments 10 times on each of the 30 datasets. We draw the mean results (precision, recall and F-score) of the three algorithms in Figure 5. The local structure learning algorithm PC-Selects achieves similar performance as the global structure learning algorithm in $\text{Adj}(W)$ and $\text{Adj}(Y)$, indicating that adjustment set discovery in data by local search is reliable.

4.2 Evaluating DICE with real-world datasets

We evaluate the effectiveness of DICE on three real-world datasets: Jobs training (Jobs for short) [18], IHDP [12] and Twins [2]. A brief description of the datasets is shown in Table 3.

| Name  | #treated | #control | #samples | #variables |
|-------|----------|----------|----------|------------|
| Jobs  | 297      | 2915     | 3214     | 9          |
| IHDP  | 139      | 608      | 747      | 24         |
| Twins | 3275     | 1546     | 4821     | 40         |

Jobs consists of the original LaLonde dataset (297 treated samples and 425 control samples) [18] and the Panel Study of Income Dynamics (PSID) observational group (2490 control samples) [14]. Each sample has 9 variables, including 7 pretreatment variables (which are age in years, schooling in years, indicators for black and
Table 4. Results on Jobs. Bias (%) is the relative error comparing to the ground-truth causal effect. The estimation with the lowest bias is highlighted in each category.

| Method  | CE   | Bias (%) | Remarks          |
|---------|------|----------|------------------|
| LV-IDA  | -0.007 | 72%       | The best of 36 estimations |
| DICE    | -0.027 | 8.0%      | The best of 64 estimations |
| DICE    | -0.039 | 59.2%     | The most probable estimate |
| PSM     | -0.018 | 56.7%     |                  |
| CBPS    | -0.010 | 59.4%     |                  |
| CFR     | -0.011 | 56.3%     |                  |
| LASSO   | -0.016 | 36.2%     |                  |
| BART    | -0.010 | 41.8%     |                  |
| TMLE    | -0.016 | 32.9%     |                  |
| CF      | -0.017 | 32.6%     | The best of 30 forests |

Table 5. Results on IHDP w.r.t. CE and Bias.

| Method  | CE | Bias (%) | Remarks          |
|---------|----|----------|------------------|
| LV-IDA  | 3.97 | 9.2%     | The best of 300 estimates |
| DICE    | 4.41 | 0.7%     | The best of 16 estimates |
| DICE    | 4.38 | 2.5%     | The most probable estimate |
| PSM     | 4.95 | 13.2%    |                  |
| CBPS    | 4.56 | 4.3%     |                  |
| CFR     | 4.86 | 11.0%    |                  |
| LASSO   | 5.06 | 15.7%    |                  |
| BART    | 4.69 | 7.3%     |                  |
| TMLE    | 4.86 | 11.2%    |                  |
| CF      | 4.25 | 2.9%     | The best of 30 forests |

Hispanic, marriage status, school degree, previous earnings in 1974 and 1975, and whether the 1974 earnings variable is missing), employment status with/without job training as treatment variable, and 1978 earning as the outcome variable. Because the dataset contains records of people taking part in the training only, as in [18], we estimate Average Treatment Effect on Treated (ATT) for DICE and all comparisons, against the ground truth AT which is $886$.

IHDP is related to the Infant and Health and Development Program (IHDP) on low-birth-weight premature infants [12]. IHDP has 24 pretreatment variables and we follow the method in [12] to generate the simulated outcome with the ground-truth of 4.38 for the causal effect (CE) by the `R` package `npec`.

Table 6. Results on Twins w.r.t. CE and Bias.

| Method  | CE | Bias (%) | Remarks          |
|---------|----|----------|------------------|
| LV-IDA  | -0.007 | 72%       | The best of 36 estimations |
| DICE    | -0.027 | 8.0%      | The best of 64 estimations |
| DICE    | -0.039 | 59.2%     | The most probable estimate |
| PSM     | -0.018 | 56.7%     |                  |
| CBPS    | -0.010 | 59.4%     |                  |
| CFR     | -0.011 | 56.3%     |                  |
| LASSO   | -0.016 | 36.2%     |                  |
| BART    | -0.010 | 41.8%     |                  |
| TMLE    | -0.016 | 32.9%     |                  |
| CF      | -0.017 | 32.6%     | The best of 30 forests |

The Twins dataset consists of samples from twin births in the USA between 1989 and 1991 with the birth weight less than 2000g. [2] We eliminate records with missing values and have 4821 twin pairs left. The treatment variable is birth weight: $W=1$ for a baby who is heavier in a twin; $W=0$ otherwise. The mortality after one year is the true outcome for each twin, and the ground-truth causal effect is -0.025. To simulate an observational study, according to [21], we use the follow-
Twin dataset, the largest absolute bias is only 0.015, and hence we consider that the estimates by all methods are similar and good.

To show that those comparison methods do not work in a dataset when the unconfoundedness assumption is not satisfied, we have generated datasets with $M$-bias. The performance of some methods deteriorates greatly. Implementations of other methods do not work with datasets with more than 200K records. We do not show the results in this paper due to the space limitation.

4.3 Efficiency evaluation of DICE

As PC-Select is a major contributor to the complexity of DICE, in this section, we firstly evaluate the efficiency of PC-Select, and then the overall efficiency of DICE. The computations were performed on a PC with 2.6 GHz Intel Core i7 and 16GB of memory.

Efficiency evaluation of the structure learning algorithms. We record the running time of the structure learning algorithms on the synthetic datasets introduced in Section 4.1 and draw the mean running time in Figure 6. PC-Select is faster than FCI and RFCI, and comparing to FCI and RFCI, PC-Select scales with the number of variables and samples very well.

Figure 6. The runtime on two groups of synthetic datasets w.r.t. the number of variables and samples.

Experiments on ten standard benchmark BNs. To evaluate the efficiency of DICE, we use the datasets generated from the ten benchmark BNs from the BN repository\(^5\): SACHS, CHILD, INSURANCE, WATER, ALARM, BARLEY, HAILFINDER, HEPAR II, WIN95 and ANDES. With each BN, we choose a variable without child nodes as the outcome variable and then select one of its parents without child nodes (except the outcome) as the treatment variable. We generate ten synthetic datasets from the ten BNs with 5000 samples each using the \texttt{R} package \texttt{bnlearn}\(^2\). Then we hide 5% variables (only one variable is hidden for the small BNs SACHS and CHILD) which lie on a non-causal path between the treatment and outcome variables. The same parameter settings are used for DICE and LV-IDA as in Subsection 4.2.

Results. As shown in Figure 7, DICE is faster than LV-IDA across all the 10 datasets. Especially, LV-IDA did not return results for the 5 datasets generated using the larger BNs within two hours, while DICE completed in seconds (at most 50+ seconds) for all datasets.

Figure 7. The runtime on ten BNs. Note: LV-IDA did not return results in two hours on BARLEY, HAILFINDER, HEPAR II, WIN95 and ANDES.

5 Conclusion

Causal query in data is an important means for evidence-based decision making, without relying on or being restricted by domain knowledge. However, its widespread applications are hindered by the low efficiency and unsatisfactory accuracy of existing methods. In this paper, we have developed theorems which support the utilisation of efficient local causal structure discovery for causal query and assures the correctness of the result of causal query obtained with local search. Based on the theorems, we have developed DICE to query data for causal effects and the confounders potentially impacting the causal effects. The results returned by DICE provide decision makers not only valuable information on the effect of changing one variable in their systems, but also the awareness of the other factors which could influence the effect. Experiments with synthetic and real-world data have demonstrated the effectiveness and efficiency of DICE. It has been shown that DICE produces better causal effect estimation than LV-IDA, and works on datasets where LV-IDA fails. The causal effects estimated by DICE are as good as the state-of-the-art methods which use domain knowledge. In future, we will evaluate DICE and

\(^5\) http://www.bnlearn.com/bnrepository/
other methods on purely observational datasets which may include $M$-bias, and apply it to real-world biological applications.

6 Acknowledgement

This research is partially funded by ARC DP170101306 and the National Science Foundation of China (under grant 61876206). The first author is supported by China Scholarship Council.

REFERENCES

[1] Constantin F Aliferis, Alexander Statnikov, Ioannis Tsamardinos, Subramani Mani, and X neon D Koutsoukos, ‘Local causal and markov blanket induction for causal discovery and feature selection for classification part i: Algorithms and empirical evaluation’, Journal of Machine Learning Research, 11(Jan), 171–234, (2010).

[2] Douglas Almond, Kenneth Y Chay, and David S Lee, ‘The costs of low birth weight’, The Quarterly Journal of Economics, 120(3), 1031–1083, (2005).

[3] Susan Athey, Guido W Imbens, and Stefan Wager, ‘Approximate residual balancing: debiased inference of average treatment effects in high dimensions’, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 80(4), 597–623, (2018).

[4] Peter Bühlmann, Markus Kalisch, and Marloes H Maathuis, ‘Variable selection in high-dimensional linear models: partially faithful distributions and the pc-simple algorithm’, Biometrika, 97(2), 261–278, (2010).

[5] Ruichu Cai, Feng Xie, Clark Glymour, Zhifeng Hao, and Kun Zhang, ‘Triad constraints for learning causal structure of latent variables’, in NeurIPS, pp. 12863–12872, (2019).

[6] Ruichu Cai, Zhenjie Zhang, and Zhifeng Hao, ‘Causal gene identification using combinatorial v-structure search’, Neural Networks, 43, 63–71, (2013).

[7] Hugh A Chapman, Edward I George, Robert E McCulloch, et al., ‘Bart: Bayesian additive regression trees’, The Annals of Applied Statistics, 4(1), 266–298, (2010).

[8] Diego Colombo, Marloes H Maathuis, Markus Kalisch, and Thomas S Richardson, ‘Learning high-dimensional directed acyclic graphs with latent and selection variables’, The Annals of Statistics, 39(4), 1406–1413, (2011).

[9] Xavier De Luna, Ingeborg Waernbaum, and Thomas S Richardson, ‘Covariate selection for the nonparametric estimation of an average treatment effect’, Biometrika, 99(4), 861–875, (2011).

[10] Doris Entner, Patrik Hoyer, and Peter Spirtes, ‘Data-driven covariate selection for nonparametric estimation of causal effects’, in Artificial Intelligence and Statistics, pp. 256–264, (2013).

[11] Jenny Haggström, ‘Data-driven confounder selection via markov and bayesian networks’, Biometrics, 74(2), 389–398, (2018).

[12] Jennifer L Hill, ‘Bayesian nonparametric modeling for causal inference’, J. Comput. Graph. Statist., 20(1), 217–240, (2011).

[13] Aniti Hyytinm, Frederick Eberhardt, and Matti Jürvissalo, ‘Do–calculus when the true graph is unknown’, in IAI, pp. 395–404, (2015).

[14] Kosuke Imai and Marc Ratkovic, ‘Covariate balancing propensity score’, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 76(1), 243–263, (2014).

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

[16] Fredrik Johansson, Uri Shalit, and David Sontag, ‘Learning representations for counterfactual inference’, in ICML, pp. 3020–3029, (2016).

[17] Markus Kalisch, Martin Mächler, Diego Colombo, Marloes H Maathuis, Peter Bühlmann, et al., ‘Causal inference using graphical models with the r package pgcai’, Journal of Statistical Software, 47(11), 1–26, (2012).

[18] Robert J LaLonde, ‘Evaluating the econometric evaluations of train– Thuc Duy Le, Lin Liu, Anna Tsykin, Gregory J Goodall, Bing Liu, Bing-Yu Sun, and Jiuyong Li, ‘Inferring microrna–mrna causal regulatory relationships from expression data’, Bioinformatics, 29(6), 765–771, (2013).

[19] Jiuyong Li, Thuc Duy Le, Lin Liu, Jixue Liu, Zhou Jin, Bingyu Sun, and Saisai Ma, ‘From observational studies to causal rule mining’, ACM TIST, 7(2), 14, (2016).

[20] Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling, ‘Causal effect inference with deep latent-variable models’, in NeurIPS, pp. 6446–6456, (2017).

[21] Marloes H Maathuis, Markus Kalisch, Peter Bühlmann, et al., ‘A generalized back-door criterion’, The Annals of Statistics, 43(3), 1060–1088, (2015).

[22] Marloes H Maathuis, Markus Kalisch, Peter Bühlmann, et al., ‘Estimating high-dimensional intervention effects from observational data’, The Annals of Statistics, 37(6A), 3133–3164, (2009).

[23] Daniel Malinsky and Peter Spirtes, ‘Estimating causal effects with ancestral graph markov models’, in Conference on Probabilistic Graphical Models, pp. 299–309, (2016).

[24] Paul R Rosenbaum and Donald B Rubin, ‘The central role of the propensity score in observational studies for causal effects’, Biometrika, 70(1), 41–55, (1983).

[25] Marco Scutari, ‘Learning bayesian networks with the bnlearn r package’, arXiv preprint arXiv:0908.3817, (2009).

[26] Jasjeet S Sekhon, ‘Multivariate and propensity score matching software with automated balance optimization: the matching package for r’, Journal of Statistical Software. Forthcoming, (2008).

[27] Uri Shalit, Fredrik D Johansson, and David Sontag, ‘Estimating individual treatment effect: generalization bounds and algorithms’, in Proceedings of the 34th ICML-Volume 70, pp. 3076–3085, (2017).

[28] Peter Spirtes, Clark N Glymour, Richard Scheines, David Heckerman, Christopher Meek, Gregory Cooper, and Thomas Richardson, Causation, prediction, and search, MIT press, 2000.

[29] Robert Tibshirani, ‘Regression shrinkage and selection via the lasso’, Journal of the Royal Statistical Society: Series B (Methodological), 58(1), 267–288, (1996).

[30] Mark J Van Der Laan and Daniel Rubin, ‘Targeted maximum likelihood learning’, The International Journal of Biostatistics, 2(1), (2006).

[31] Benito van der Zander, Maciej Liskiewicz, and Johannes Texton, ‘Constructing separators and adjustment sets in ancestral graphs.’, in C的 IAJ, pp. 11–24, (2014).

[32] Tyler J VanderWeele and Ilya Shpitser, ‘A new criterion for confounder selection’, Biometrics, 67(4), 1406–1413, (2011).

[33] Stefan Wager and Susan Athey, ‘Estimation and inference of heterogeneous treatment effects using random forests’, Journal of the American Statistical Association, 113(523), 1228–1242, (2018).

[34] Kui Yu, Lin Liu, and Jiuyong Li, ‘Learning markov blankets from multiple interventional data sets’, IEEE Trans. Neural Netw. Learn. Syst., 9, (2019).

[35] Kui Yu, Lin Liu, Jiuyong Li, and Huanhuan Chen, ‘Mining markov blankets without causal sufficiency’, IEEE Trans. Neural Netw. Learn. Syst., 29(12), 6333–6347, (2018).

[36] Kui Yu, Lin Liu, Jiuyong Li, Wei Ding, and Thuc Duy Le, ‘Multi-source causal feature selection’, IEEE Transactions on Pattern Analysis and Machine Intelligence, (2019).

[37] Jiji Zhang, ‘Causal reasoning with ancestral graphs’, Journal of Machine Learning Research, 9(Jul), 1437–1474, (2008).